DOCTOR OF PHILOSOPHY

Analysis of variation in protein domain families and interfaces

Marques Madeira, Fabio Manuel

Award date:
2016

ANALYSIS OF VARIATION IN PROTEIN DOMAIN FAMILIES AND INTERFACES

By

Fábio Manuel Marques Madeira

Thesis submitted for the degree of Doctor of Philosophy

at

University of Dundee

Dundee, United Kingdom

September 2016

Contents

List of Figures

List of Tables

List of Abbreviations

1kGP	1,000 Genomes Project
3D	Three-dimensional
API	Application Program Interface
AL	Alignment Length
AsymUnit	Asymmetric Unit
BioLiP	Ligand-protein Binding Database
BioUnit	Biological Unit
bp	Base Pair
C.I.	Confidence Interval
Cα	Alpha-carbon
CATH	Class, Architecture, Topological Motif and Homologous Superfamily
ClinVar	Archive of Interpretations of Clinically Relevant Variants
COSMIC	Catalogue of Somatic Mutations in Cancer
CryoEM	Cryo-electron Microscopy
Da	Dalton
dbSNP	The Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic Acid
DP	Dynamic Programming
EC	Enzyme Commission
ENCODE	The Encyclopaedia of DNA Elements Project
ESP	Exome Sequencing Project

ExAC	Exome Aggregation Consortium
FunFam	CATH Functional Family
FTP	File Transfer Protocol
GO	Gene Ontology
GRCh	Genome Reference Consortium Human Genome Build
HapMap	Haplotype Map Project
HGMD	Database of Human Gene Mutation Data
HMM	Hidden Markov Model
HTTP	Hypertext Transfer Protocol
Humsavar	UniProt's Collection of Human Polymorphisms and Disease Mutations
iRMSD	Interaction RMSD
KEGG	Kyoto Encyclopedia of Genes and Genomes
LSS	Length of the Shortest Sequence
mmCIF	Macromolecular Crystallographic Information File
MS	Mass Spectrometry
MSA	Multiple Sequence Alignment
NEP	Number of Structurally Equivalent Positions
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NMR	Nuclear Magnetic Resonance
nsSNP	Non-synonymous SNP
OMIM	Online Mendelian Inheritance in Man
OR	Odds Ratio
PDB	Protein Data Bank
PDBe	PDB in Europe
Pfam	Protein Families Database
PhenCode	Phenotypes for ENCODE
PID	Percentage Sequence Identity

PISA	Proteins, Interfaces, Structures and Assemblies
PLI	Protein-ligand Interaction
Polyphen-2	Polymorphism Phenotyping Predictor version 2
PPI	Protein-protein Interaction
PQS	Protein Quaternary Structure Server
ProIntVar	Protein Interactions and Variants
PTM	Post-translational Modification
REST	Representational State Transfer
RESTful	A Web Service Based on REST
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
RSA	Relative Solvent Accessibility
$\boldsymbol{Sc}$	STAMP Structural Similarity Score
SC	CATH Structural Cluster
SCR	Structurally Conserved Region
SCOP	Structural Classification of Proteins
SF	Structural Fragment
SIFT	Sorting Tolerant From Intolerant Predictor
SIFTS	Structure Integration with Function, Taxonomy and Sequence
SMCRA	'Structure, Model, Chain, Residue, Atom' Data Model
SNP	Single Nucleotide Polymorphism
SPF	Superfamily
SSE	Secondary Structure Element
STAMP	Structural Alignment of Multiple Proteins
UniProt	The Universal Protein Resource
UniProtKB	UniProt KnowledgeBase

Acknowledgements

I would like to thank Geoff Barton for all his guidance and advice, and for sharing many invaluable memories and experiences. I would like to thank Ludwig Krippahl and Carol MacKintosh for mentoring me at an earlier stage. In addition, thanks to everyone in the Division of Computational Biology, particularly, Thiago Britto-Borges, Nancy Giang, Stuart MacGowan, Michele Tinti, Jim Procter, Chris Cole, Nick Schurch and Mungo Carstairs, for all the interesting discussions and fun. I would like to thank my dearest, Daniela Penas, for her endless love and encouragement. I would like to thank my parents and brother, as well as my close family and friends, for their comfort and support. Finally, I would like to thank the Wellcome Trust for funding my Ph.D. project, and the School of Life Sciences for empowering my career.

To my Father

STATEMENT

School of Life Sciences, University of Dundee

I certify that Fábio Manuel Marques Madeira has satisfied all the terms and conditions of the relevant Ordinance and Regulations to qualify in submitting this thesis, entitled 'Analysis of variation in protein domain families and interfaces', in application for the degree of Doctor of Philosophy.

Date: September 2016

Supervisor: _____
Prof. Geoffrey J. Barton

DECLARATION

School of Life Sciences, University of Dundee

I hereby declare that the work described in this thesis, entitled 'Analysis of variation in protein domain families and interfaces', is my own; that I am the author of this thesis; that it has not previously been put forward in submission for any other degree or qualification; and that I have consulted references herein.

Date: September 2016

Candidate: _____
Fábio Manuel Marques Madeira

Abstract

There are currently more than 100 million single nucleotide polymorphisms (SNPs), which are the most frequent and basic type of genetic variation. The availability of over 110,000 three-dimensional protein structures allows the structural context of many SNPs to be examined in atomic detail. Interfaces are essential sites for protein function and adaptation, and key in a majority of biological processes. A computational framework, ProIntVar, was developed for mapping SNPs onto the structures to allow the features of variation at domain-domain and domain-ligand interfaces to be studied. ProIntVar allows the systematic analysis of genetic variation in protein structure interaction surfaces by integrating structural and sequencing data from several biological databases and resources. Protein domains and variants were analysed in the context of structural clusters (SCs) and functional families (FunFams), which are derived from structurally and functionally related protein domains in Superfamilies classified in CATH (Class, Architecture, Topology, Homologous Superfamily). Multiple structural alignments were generated by STAMP for each CATH SC and FunFam, using an improved protocol that leads to improvement of the structural superimposition and resulting structure-based alignments. The structural alignments were extended with similar protein sequences by HMM-based

sequence search. These sequences are believed to be structurally/functionally homologous and thus a rich source of novel insight into the structural context and potential consequences of a vast number of genetic variants. The characterisation of both disease-associated and non-disease germline variants, as well as somatic variation, was performed. The analysis of non-synonymous SNPs (nsSNPs) was stratified by annotation and potential consequence and focused particularly on domain-domain and domain-ligand interaction interfaces. Domain interactions were screened and further classified by mode of interaction as clustered by iRMSD (interaction root-mean-square deviation). The results corroborate previous observations that pathogenic mutations are enriched at key sites, such as structurally conserved domain-ligands interfaces and the protein core. Examination of genetic variation at such hot-spots in the context of domain families helps to infer which variants are more likely to affect protein activity and function in a broader evolutionary sense. The most drastic features shared by pathogenic variants were identified to prioritise the analysis of nsSNPs currently thought to be neutral, but potentially disruptive.

Chapter 1

Introduction

This Chapter reviews the state of the art in the two main areas covered in this Thesis: 1) protein structures and interactions, and 2) genetic variation. This is followed by a brief overview of the approaches undertaken to improve upon current methods of analysis of genetic variation at protein domain interfaces.

1.1 Background

The central dogma of molecular biology (Crick, 1970) unveils the sequential transfer of information from the genetic material to its translated products known as proteins. DNA is replicated, then transcribed into RNA. After a series of processing steps, which includes splicing, the RNA molecule is translated into proteins. Proteins are considered the most important machinery of the cell (Gutteridge and Thornton, 2005; Scaiewicz and Levitt, 2015) and perform nearly every function required for life. Proteins participate extensively in interactions, which are fundamental to virtually all biochemical processes (Alves et al., 2002; Nooren and Thornton, 2003b; Russell

et al., 2004; Schuster-Böckler and Bateman, 2008; Marsh and Teichmann, 2015). Genetic variation and the accumulation of sequence variation is an important source of variability within and across populations, and an important driver of evolution inherent to all living organisms (Nei et al., 2010; Simonti and Capra, 2015). The study of sequence variation is key to the identification of random mutations linked to impairment of protein function, through the disruption of interaction interfaces and contact networks (Sunyaev et al., 2001; Wang and Moult, 2001; Ramensky et al., 2002; Yue and Moult, 2006; Stefl et al., 2013).

There has been a tremendous increase in the amount of biological data being generated over the last decades. This includes a dramatic increase in the availability of genomics, proteomics, and structural data, among others. According to data from the Genomes Online Database (GOLD) database, there were only 350 active genome sequencing projects in 1997 (Bernal et al., 2001), which had increased to 11,472 by 2011 (Pagani et al., 2012). The genomes of over 11,006 organisms have now been fully sequenced in comparison to only 48 in 1997. Associated with these genetics studies is the identification of sequence variations (genetic variants). dbSNP (Sherry et al., 1999) accounted for a mere 4,713 unique variants in 1999. This number increased rapidly and accounted for more than 60 million genetic variants in 2013 (Chen et al., 2010).

Another major data source is proteomics, where the number of public datasets and experiments in the Proteomics Identifications Database (PRIDE) increased dramatically to over 25,853 in 2012, accounting for more than 10 million identified proteins in over 320 species (Vizcaíno et al., 2013). The emergence of proteomics-based

techniques has also lead to a remarkable increase in the availability of protein-protein interactions data. The Molecular Interaction database (IntAct) contained around 2,200 binary interactions in 2004 (Hermjakob et al., 2004), which increased to 430,134 interactions in 2013 (Orchard et al., 2014). In a similar trend, there were 5,009 protein structures deposited in the Protein Data Bank (PDB) in 1996 (Westbrook et al., 2002), which increased to over 100,000 structures in 2014 (Gutmanas et al., 2014).

The fields of biological sciences that traditionally studied small datasets related to their particular area of expertise have also evolved in response to this data deluge. The huge demand for interpretation and analysis of these vast amounts of data is being managed by the emerging field of bioinformatics (Bayat, 2002; Ouzounis and Valencia, 2003). Bioinformatics is an interdisciplinary field that draws from biological sciences (biology, biochemistry, genetics); medicine (populations, cells, cancer, etc.), computer science (software development, database technology, artificial intelligence, etc.); as well as physics and maths (mathematical modelling, statistics, etc.) (Roos, 2001; Bayat, 2002; Chicurel, 2002; Larranaga, 2006; Hamelryck, 2009; Berman et al., 2007; Pruitt et al., 2003; Xia and Levitt, 2004; Aloy and Russell, 2006; Taylor, 2007; Maglott et al., 2011). Major challenges in biological sciences and bioinformatics persist regarding interpretation and functional annotation of the large datasets generated. The functional annotation of protein interactions and the potential consequences of genetic variation are among those that are addressed in this Thesis.

1.2 Proteins

Proteins are linear polymers of amino acids. The distinct sequence of amino acid residues in a polypeptide chain determines the three-dimensional (3D) structure of the folded protein. The amino acid sequence is referred to as the primary structure. Secondary structure is the local conformation of the polypeptide chain, which is stabilised by hydrogen bonding and the properties of peptide bonds between amino acid residues. The dominant secondary structure elements (SSEs) in proteins are α-helices and β-strands. These regular structures are distributed within irregular, generally ordered loops, turns or coils, as well as disordered loop regions, which are referred to as random coils. Loop regions are often located at the surface of the protein, and in addition to simply serving as transitions between regular structures, often harbour active sites in enzymes (Worth et al., 2009).

The arrangement of the secondary structure elements in space is referred to as the tertiary structure. The tertiary structure is locally governed by the interactions between amino acid residue side chains. Globally, and most importantly, the tertiary structure results from the hydrophobic effect (Kumar and Nussinov, 2002), where residues with hydrophobic side chains are packed into the core of the protein, away from the solvent. In the hydrophobic core of the protein, the polarity of the polypeptide backbone is neutralised by hydrogen bonding in SSEs. Buried polar residues form hydrogen bonds with other polar residues or the polypeptide backbone and in some cases with integral water molecules contained inside the protein. Charged residues in the hydrophobic core form ionic interactions with residues of opposite charge. Disulphide bonds formed between Cys residues are the only type of covalent

interaction formed between amino acid residues. Disulphide bonds further stabilise tertiary structure, but they are not found in all proteins. The structure-stabilising role of interactions between buried residues has been hypothesised long ago, but the contribution of surface residues to protein stability was also observed (Strickler et al., 2006).

Quaternary structure refers to the organisation of monomers in multi-subunit proteins, stabilised by similar interactions as the secondary and tertiary structures. The interfaces between monomers are often thought to be hydrophobic and thus resemble the cores of globular proteins. This has been shown to be true for homodimeric proteins, because they rarely occur as monomers, and hence their interaction surfaces are permanently buried within the protein-protein complex. Heteromers, instead, often occur and function as monomers in solution and thus the interfaces of transient complexes exhibit a more hydrophilic nature (Janin and Chothia, 1990; Jones and Thornton, 1996, 1997). These interfaces could not be as hydrophobic as those of homodimers because a large solvent-exposed hydrophobic area on the protein would be energetically unfavourable (Jones and Thornton, 1996). The properties of protein interactions are further explored in Section 1.6.

1.3 Protein sequence

The sequences of over 40 million proteins have been determined and collected in biological sequence databases (Cochrane et al., 2010). The Universal Protein Resource (UniProt) (The UniProt Consortium, 2015) is one of the largest protein information databases, and it consists of the UniProt Knowledgebase (UniProtKB), the UniProt

Reference Clusters (UniRef), and the UniProt Archive (UniParc). The UniProtKB database is the central unit of the UniProt database, and is a collection of sequence and functional information on proteins, with a comprehensive selection of annotations for each protein. It contains detailed information about gene products, partial lists of mutations and links to other data sources. The UniProtKB can be split into two essential parts: the UniProtKB/SwissProt, which contains fully manually annotated records with additional information taken from the scientific literature and manually evaluated computational analysis, and the UniProtKB/TrEMBL, which consists of automatically generated entries that await full manual annotation.

A great effort has been made over the last decades to annotate the protein sequences with relevant and accurate information. Functional annotation involves several properties of protein function including post-translational modifications (PTMs), expression levels, transcriptome, biomolecular interactions and cellular location. Complementary sequence annotation features are provided by many databases such as the Gene Ontology (GO) (Harris et al., 2004), Enzyme Commission (EC) (Bairoch, 2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), and Protein Families (Pfam) (Bateman et al., 2004) databases.

1.4 Protein structure

The major public repository for structural data is the worldwide Protein Data Bank (PDB) central archive of macromolecular structural data (Westbrook et al., 2002; Berman et al., 2003, 2007). The PDB is an archival database for macromolecular structures established in 1971 by Brookhaven National Laboratory, New York,

as a public domain repository. The first version of the PDB contained only seven structures. There has been a dramatic increase in the number of deposited structures since then, and projects such as the 'Structural Genomics Initiatives', aimed at solving the three-dimensional structure of selected protein families and protein complexes, known to be potential drug targets or important players in key pathways for disease (Xie and Bourns, 2005; Arkin and Wells, 2004; Mullard, 2012; Ivanov et al., 2013). This led to a remarkable increase in the number of available protein structures in the PDB (Velankar et al., 2010), which recently surpassed the 100,000 mark. Figure 1.1 shows the cumulative increase in the number of macromolecular structures available in the PDB over the last 40 years.

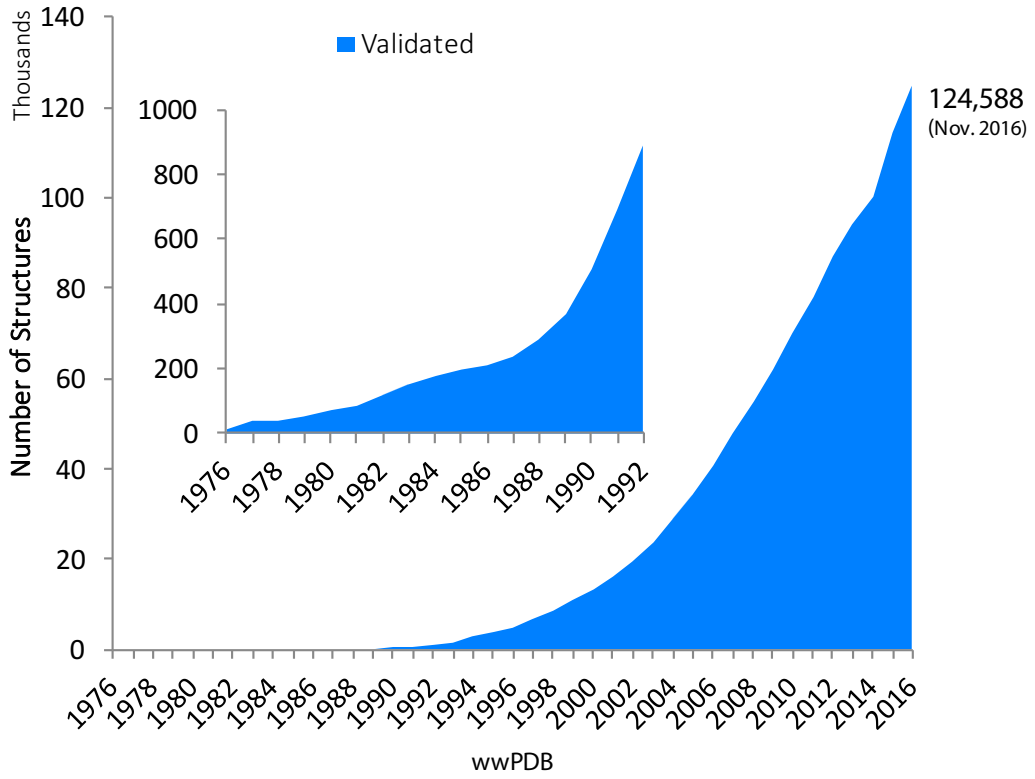


Figure 1.1: The exponential growth in the number of macromolecular structures deposited in the worldwide Protein Data Bank (wwPDB). Data obtained from `http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html` as accessed in November 2016.

1.4.1 Experimental methods

The most commonly used experimental techniques for solving 3D structures are X-ray crystallography (Gilliland and Ladner, 1996; Carpenter et al., 2008; Wlodawer et al., 2007; Shi, 2014), Nuclear Magnetic Resonance (NMR) (Luca et al., 2004; Englander and Mayne, 1992; Kwan et al., 2011) and Cryo-(transmission) Electron Microscopy (cryoEM) (Saibil, 2000; Nickell et al., 2006; Jonic and Vénien-Bryan, 2009). X-ray crystallography can obtain atomic level resolution (1.5 Å or lower), whereas cryoEM usually produces images of lower resolution ($\geq$3.0 Å). Each method determines the 3D structure of proteins exploring different protein properties and experimental conditions. This means that not all structures can be determined by any one of the methods, and it is now common practice to use these complementary techniques to solve the structure of proteins. This is especially true for bigger protein complexes (Lander et al., 2012; Lengyel et al., 2014; Hashem et al., 2013; Jackson et al., 2015; Schröder, 2015). There are additional techniques that can provide insight into the shape of macromolecules, such as Small Angle X-ray Scattering (SAXS) (Putnam et al., 2007) and Fluorescence Resonance Energy Transfer (FRET) (Lilley and Wilson, 2000; Schuler and Eaton, 2008). The techniques for purification and preparation of the target protein samples, as well as the experimental techniques themselves, are under constant active development. Among the most prominent emerging techniques is MicroED that uses protein microcrystals (Shi et al., 2013; Nannenga and Gonen, 2014).

1.4.1.1 X-ray crystallography

X-ray crystallography provides detailed atomic models and produces the highest resolution 3D structures, in comparison to other 3D structure determination methods (Shi, 2014). Of the more than 100,000 entries in the PDB, about 90% were determined by this technique. It has also contributed to more than a dozen Nobel prizes over the years.

Figure 1.2 illustrates the process of solving the crystal structure of macromolecules by X-ray diffraction (Gilliland and Ladner, 1996). Individual steps in structure determination include: 1) getting a protein crystal suitable for the experiment, with adequate quality and size; 2) obtaining the diffraction pattern with the appropriate wavelength; 3) evaluating the diffraction pattern to get the lattice parameters (unit cell), symmetry (space group) and diffraction intensities; 4) solving the electron density Fourier transform equation, obtaining any information about the phases of the diffracted beams (phase problem), which is a key point for the structural resolution; 5) building an initial structural model that fits the electron density by completing the model locating the remaining atomic positions; 6) iteratively refining the model, re-adjusting all atomic positions to get the best possible fit between the calculated diffraction pattern and the experimental diffraction pattern, and finally; 7) validating the structural model obtained. Among these steps, crystallisation and solving the phase problem constitute the hardest to resolve.

Crystallisation has remained a time-consuming trial and error approach that often involves screening of many thousands of crystallisation conditions and testing alternative protein constructs (Gilliland and Ladner, 1996). The proteins within a

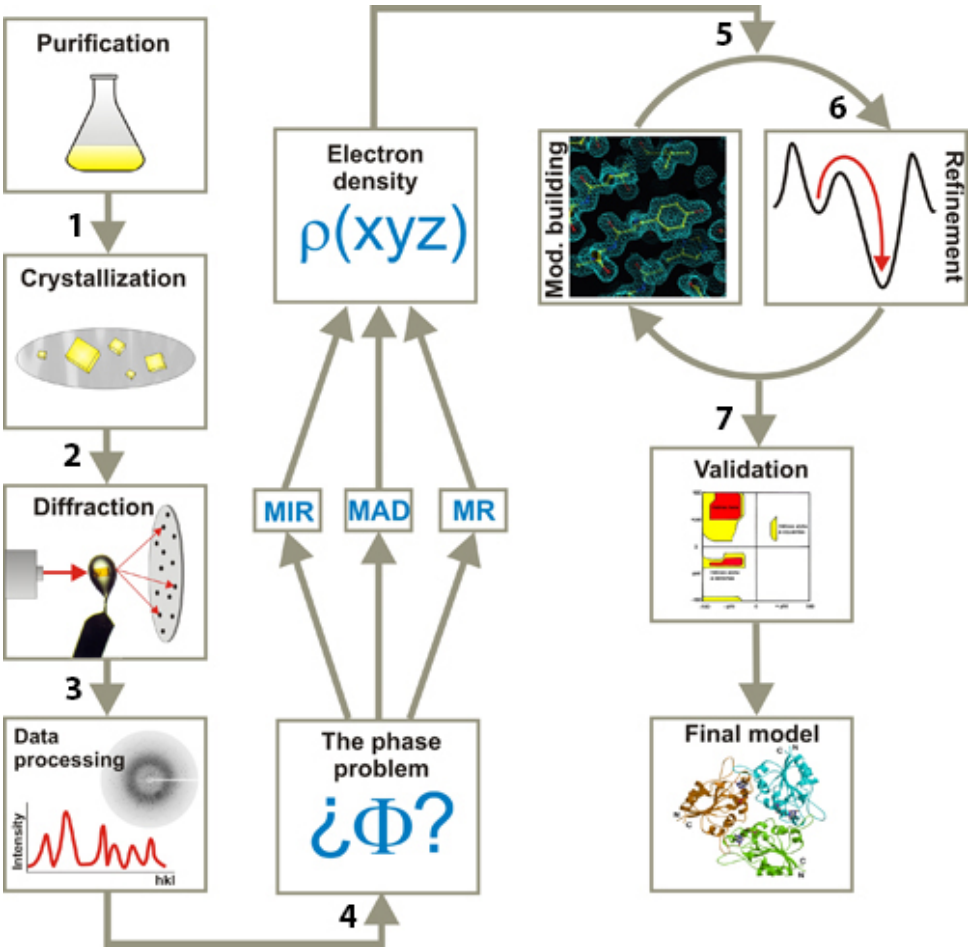


Figure 1.2: General diagram illustrating the process of solving the structure of macromolecules by X-ray crystallography. The process broadly consists of the following steps: 1) getting a protein crystal with adequate quality and size; 2) obtaining the diffraction pattern; 3) evaluating the diffraction pattern; 4) solving the electron density Fourier transform; 5) building an initial structural model; 6) refining the model; and finally; 7) validating the structural model obtained. Figure adapted from `http://www.xtal.iqfr.csic.es/Cristalografia/parte_07-en.html`.

crystal are arranged in the same orientation, which leads to interference between the diffracted waves and adds up in phase increasing the signal to a measurable level. The diffraction pattern consists of spots of various intensities arranged on a 3D lattice in space which are then used to derived the position of atoms in the 3D space.

To determine the structure of the protein both the amplitude and a phase must

be determined. The amplitude is reasonably easy to find as it is directly proportional to the brightness of the spots. To solve the phase problem, several approaches are typically used, which involve using relationships between diffraction patterns of similar crystals and different wavelengths of the same crystal (Figure 1.2). These include: Molecular Replacement (MR), when the coordinates of a similar protein are already available; Multiple Isomorphous Replacement (MIR), for when such model is not available, which relies on the addition of heavy atoms into specific sites within the unit cell without perturbing the crystal lattice; and lastly Multiwavelength Anomalous Dispersion (MAD), which is an effective approach that relies on the measurement difference produced by the introduction of one or more anomalously scattering molecules such as selenomethionines. Given the amplitude and phase of each position in reciprocal space, the electron density can be determined by applying the Fourier transform equation. The electron density map is then interpreted in terms of a set of atomic coordinates, usually starting by the fitting of the protein backbone. Once a preliminary model of the protein structure is obtained, iterative re-phasing, model fitting and refinement, is performed until a final model that leads to an agreement between the model and the data is generated (Wlodawer et al., 2007).

1.4.2 Structural data

Although historically macromolecule structures were stored in the PDB-file format, a new exchange format has been adopted. Macromolecular Crystallographic Information File (mmCIF) [`http://mmcif.pdb.org/`] is based on the Self Defining Text Archival and Retrieval (STAR) (Hall, 1991) format, and an expansion of the

Crystallographic Information File (CIF) format, which is the International Union of Crystallography (IUCr) standard for representing small molecules. The PDB format could not adequately accommodate a large amount of data associated with a single macromolecular complex structure (Bourne et al., 1997). All structural data in mmCIF consists of categories of information represented as tables and keyword-value pairs. Each data item is defined in the PDBx Exchange Data Dictionary, and chemical descriptions of all monomers and ligands in the structures are provided in the PDB Chemical Component Dictionary and Biologically Interesting molecule Reference Dictionary (BIRD).

1.4.2.1 Quality of the protein models

Several validation metrics have been developed to assess the quality of protein structure models (Brunger, 1992; Domagalski et al., 2014). Many quantitative measures have been devised that correlate with model quality, including resolution, R-factor and R-free, Ramachandran distribution, distribution of deviations from ideal geometry, Molprobity's clash-score (Hintze et al., 2016), among others. Nevertheless, no single parameter is sufficient to conclusively determine whether a given structure is of high or low quality, and quality depends on the context. All structures deposited and organised in the PDB are submitted through a validation server (Gore et al., 2012) in order to assess how well the protein model fits the data. Additionally, every structure is compared to other structures in the PDB using a variety of established validation methods (Gore et al., 2012; Velankar et al., 2013). Validation reports are provided for all structures covering experimental information, data and refinement statistics, as well as validation information. The main summary metrics currently

provided for each structure are R-free, clash-score, Ramachandran and side-chain outliers, as well as Real Space R-value Z-score (RSRZ) outliers.

The majority of the structures in the PDB have resolutions ranging from 1.6 Å to 2.8 Å, and it is estimated that at this resolution they are likely to have errors associated with them (Jones et al., 1991; Acharya and Lloyd, 2005). These errors include the atomic positions, the planarity of peptide bonds, bond lengths, bond angles and torsion angles. Most of these errors occur in regions of high mobility or multiple conformations (Hintze et al., 2016).

1.4.2.2 Problems with the structural data

Although the information found in PDB structures are clearly invaluable, high-throughput analysis of the structures is difficult due to the heterogeneous and inconsistent nature of the data. Inconsistencies result from a variety of sources: crystal structures with multiple-occupancy atoms; inconsistent presence of water molecules; inconsistent presence of hydrogen atoms; atoms with missing electron density; structures consisting solely of Cα backbone atoms; structures containing non-standard amino acid residue types (naturally-occurring or engineered modifications forming part of the polypeptide backbone); among others. Additionally, many of the problems stem from the fact that the PDB is not a relational database (Schierz et al., 2007). The situation has improved considerably through the efforts of the PDB remediation project (Henrick et al., 2007), as well as the recently introduced validation server (Gore et al., 2012).

X-ray crystallography has several limitations related to the fact that it only produces an average picture of a structure, the structure of the protein complex in

crystal form. Structures solved by NMR do not present this problem and can be used to complement the X-ray structures. Nonetheless, structures are solved in a set of conformations (ensembles) which raise the problem of choosing the best model representation of the protein (Mao et al., 2014). It is also common that large macromolecular complexes are impossible to crystallise as a whole. The usual strategy is to cut the protein into manageable pieces, which then raises another challenge of reassembling all units in the proper orientation and arrangement. Structures solved at lower resolutions (for example by cryoEM) are increasingly used as guides to help model the structure of such large complexes (Lawson et al., 2011; Patwardhan et al., 2014).

1.4.2.3 Artefacts in the structures

One way in which the field of X-ray crystallography has been developed into a high-throughput technology is to use automated methods to test multiple conditions. Usually, many parameters such as the pH, the temperature and the concentration of the additives (organic solvents and crystallisation buffers) are varied to aid the crystallisation of the molecules. Additionally, crystallisation is performed using artificially high protein concentrations, which can lead to the formation of extensive non-specific crystal packing interfaces. There are several examples of protein crystal structures which display non-biological interactions (see below Section 1.4.2.5).

Many macromolecular structures in the PDB are of partial or modified proteins (e.g. engineered mutants and modified residues, truncated proteins, synthetic constructs or chimeric constructs, and purification/expression tags). Common missing segments in the structures are usually observed for tails and loops that due to their

more disordered nature are not packed in the same exact orientation in the crystal. Discontinuities in the structures are particularly troublesome when performing structure to sequence mapping (overviewed in Section 1.4.2.6).

1.4.2.4 Ligand artefacts

Ligand artefacts are commonly observed in the PDB structures. In addition to artefact ligand molecules found in solvents and crystallisation additives, many ligands correspond to substrate analogues or inhibitors that compete with the natural molecules for binding. Many of these are not believed to be biologically meaningful. In order to address this issue, it has been common practice to establish exclusion lists of compounds considered to be crystallisation artefacts which are kept out from structure analysis (Powers et al., 2006; Schreyer and Blundell, 2009; Roy and Zhang, 2012; Elokely and Doerksen, 2013; Niedzialkowska et al., 2016). One major drawback of this approach is the fact that some ligands might be irrelevant for a particular protein structure and yet essential for others (e.g. metal ions). Since making a comprehensive list of artefact compounds is hard and might lead to additional analysis artefacts, other approaches which take into account contextual knowledge have been developed (Yang et al., 2012).

1.4.2.5 Biological assemblies

The 3D atomic coordinates that appear in PDB entries are those of the asymmetric unit (AsymUnit). The AsymUnit is a set of atoms which, when operated on by the crystallographic symmetry operations defined by the space-group, generates the complete crystal. The AsymUnit may not be the biologically relevant unit of the

structure, and so may lack some key protein-protein interactions (Jefferson et al., 2006). As such, although the AsymUnit can represent the biologically functional assembly of the protein, often it comprises only a portion of a biological molecule. A biological unit or biological assembly (BioUnit) is a macromolecular complex that is believed to represent the functional form of a molecule. Since it is non-trivial to determine the BioUnit for most PDB structures, studies of protein-protein interactions have mostly ignored the additional interactions that are potentially available.

An additional problem is that some of the interactions seen in the AsymUnit of the crystal may be artefacts of crystallisation and therefore may not be biologically relevant. The problem of reliably distinguishing biologically significant assemblies from crystal contacts is non-trivial, and several approaches have been proposed (Valdar and Thornton, 2001; Mintseris and Weng, 2003; Bahadur et al., 2004; Bernauer et al., 2008; Krissinel and Henrick, 2007). The PDB provides information on the BioUnit for each deposition. This information has been traditionally provided by depositors and more recently supplemented by supporting information from the UniProt, the Protein Quaternary Structure (PQS) (Henrick and Thornton, 1998), and Proteins, Interfaces, Structures and Assemblies at the PDB in Europe (PDBePISA or simply PISA) (Krissinel and Henrick, 2007). PDBePISA is a widely-used method, which identifies functional interactions by analysing properties of the interaction surface between chains. Once interactions have been determined, PDBePISA generates a new set of coordinates for the biologically most likely BioUnits. Given the importance of working with BioUnits, the PDBe recently started providing precomputed assemblies in mmCIF format (Velankar et al., 2016).

Figure 1.3 illustrates the process of generating BioUnits from the AsymUnit by applying biologically relevant symmetry operations and/or removing crystal packing artefacts. As illustrated in Figure 1.3 B and C, there are several instances where the asymmetric unit of a crystal contains non-biological interaction interfaces.

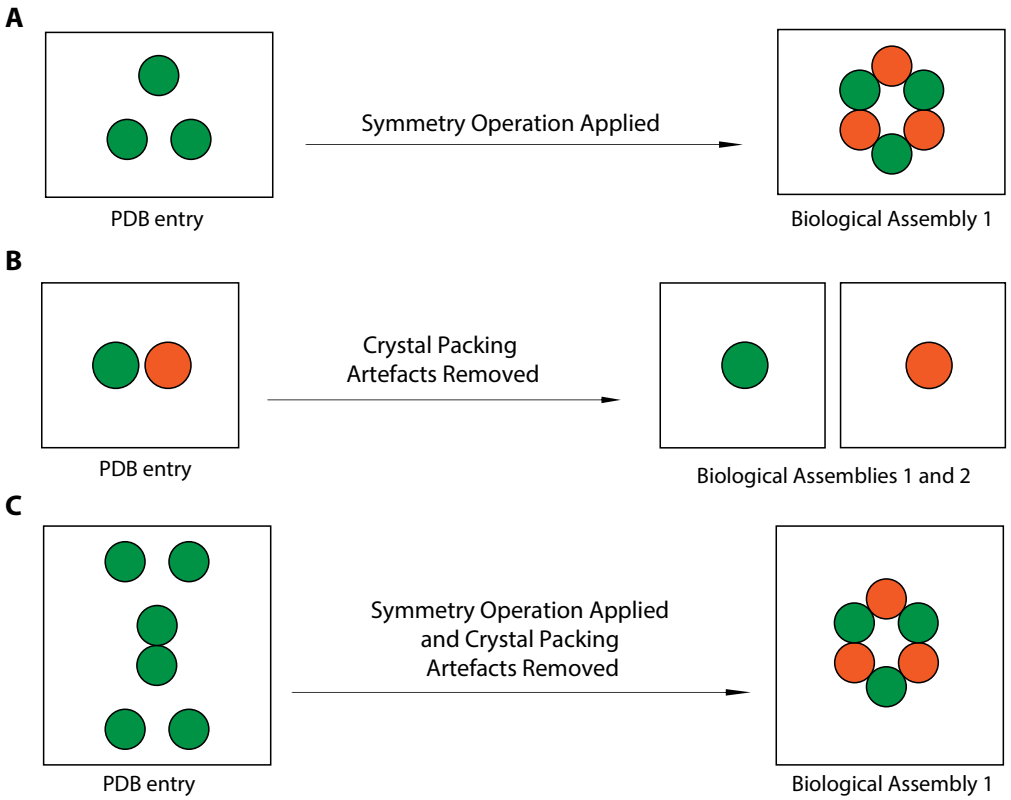


Figure 1.3: Generating biological assemblies from the asymmetric unit by applying biologically relevant symmetry operations and/or by removing crystal packing artefacts. Examples provided depict: A) generation of a biological assembly (BioUnit) by applying symmetry operations to the asymmetric unit (AsymUnit); B) generation of two BioUnits by removing crystal packing artefact interactions, and; C) generating a BioUnit by applying symmetry operations and simultaneously removing artefact crystal packing. Figure adapted from Jefferson et al., 2006.

1.4.2.6 Structure to sequence mapping

'Structure to sequence mapping' (or sequence to structure mapping) refers to the accurate mapping between a protein's sequence, as observed in a PDB structure, to a corresponding sequence record (e.g. UniProtKB entry). Ideally, this mapping

is performed at the level of individual residues observed in the structure, but it is somewhat non-trivial due to the various structural problems described in Section 1.4.2.3. The main benefits of performing structure to sequence mapping are that available sequence residue annotations can trivially be transferred to PDB structures. The Structure Integration with Function, Taxonomy and Sequence (SIFTS) resource (Velankar et al., 2013) has developed methods to align the sequences using additional annotations and manual inspection in order to provide the structure to sequence mappings. The SIFTS protocol is now routinely applied to all newly solved structures as part of the deposition and validation process in the PDB (Velankar et al., 2013).

1.5 Protein domains

Domains are considered to be the minimal and fundamental functional units of proteins and are usually associated with a specific biological role (Orengo et al., 1994; Ponting and Russell, 2002; Orengo and Thornton, 2005; Bornberg-Bauer et al., 2005). Although with a considerable diversity (Reeves et al., 2006; Taylor, 2007), most domains show similarities in sequence or structure, which reflects their origin from a common ancestor and allows them to be grouped into a hierarchy of families, superfamilies and folds. Domain analysis indicates that a limited set of thermodynamically stable folds have emerged in nature (Thornton et al., 1999; Holm and Sander, 1999). Each domain forms a compact 3D structure and often can be found independently folded (Hubbard et al., 1997; Kummerfeld and Teichmann, 2009). Many proteins consist of several structural domains, and a particular domain may

appear in a variety of different proteins (Bhaskara and Srinivasan, 2011; Han et al., 2007; Björklund et al., 2005). Fold recurrence means that few newly-solved domains are found to have new folds. This observation has been used to detect similar domains in a query structure by looking within a library of previously classified domains. Nevertheless, both the number of known folds and types of domain-domain interactions that have been identified is believed to be far from complete (Garma et al., 2012).

1.5.1 Protein structure classification

Several methods have been developed to identify and classify protein domains observed in the PDB structures (Redfern et al., 2007; Dengler et al., 2001; Lo Conte et al., 2000; Orengo et al., 2002; Holm and Sander, 1996). Class, Architecture, Topological motif and Homologous Superfamily (CATH) (Orengo et al., 2002) and Structural Classification of Proteins (SCOP) (Lo Conte et al., 2000) are the two main hierarchical domain classification systems. Domains are classified in CATH according to sequence, structural and functional similarity using both automated and manual approaches. Unlike CATH, boundary assignment and classification in SCOP is entirely achieved by manual inspection and curation. Evolutionary Classification of Protein Domains (ECOD) (Cheng et al., 2015) is a new hierarchical classification system that groups structure domains mainly according to sequence similarity.

1.5.1.1 CATH

Figure 1.4 overviews the four levels of the CATH structural classification hierarchy. Domains are initially classified according to their secondary structure content (α-helix, β-strand, mixed α and β, or very little secondary structure content). They are further classified according to Architecture, which refers to the gross arrangement of SSEs in the 3D space, independent of their connectivity. Next, Topology groups are determined by both the arrangement of SSEs and their connectivity. Finally, domains are clustered into the same Homologous Superfamily provided that clear indication of an evolutionary relationship exist. Superfamily-level domains share

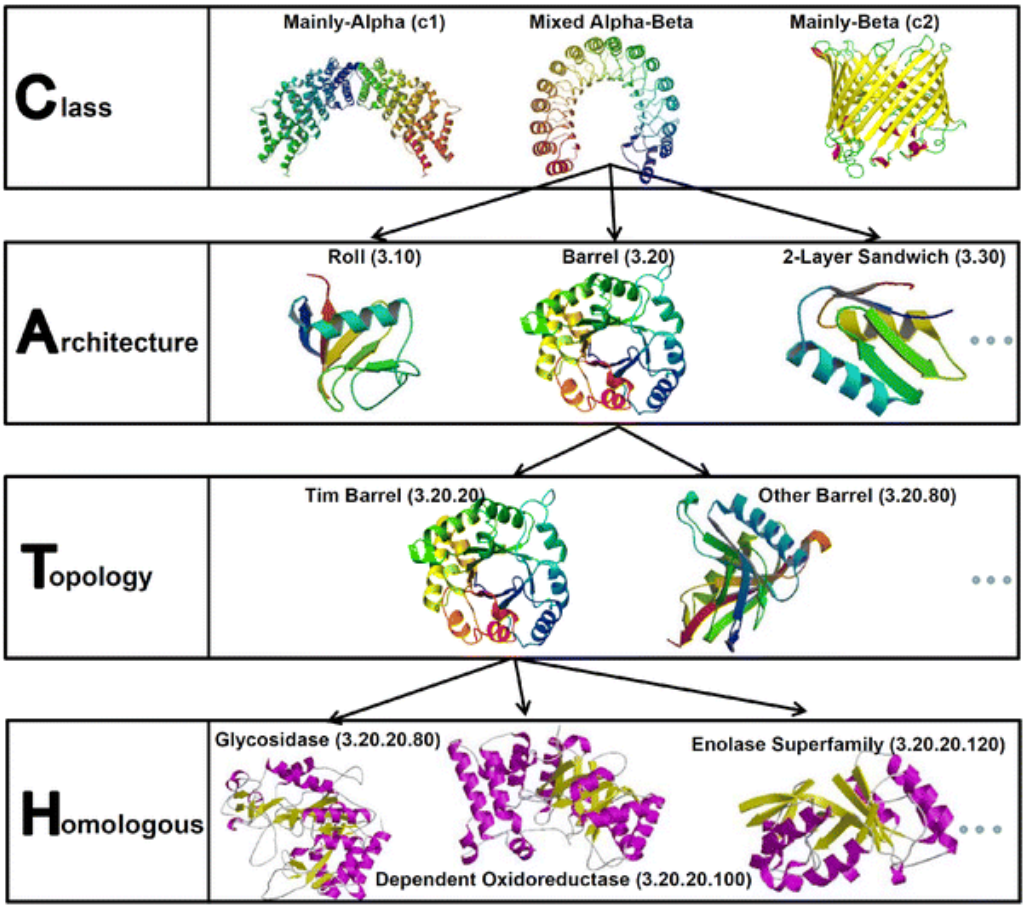


Figure 1.4: Overview of the CATH (Class, Architecture, Topology and Homologous Superfamily) structural classification hierarchy. Domains in CATH are organised according to four hierarchy levels comprising Class, Architecture, Topology and Homologous Superfamily. Figure from Maleki et al., 2013.

significant structure, sequence and/or functional similarity. Lower levels comprise subfamilies of domains at different levels of sequence similarity (45, 60, 95, 100 percentage sequence identity (PID)). Domains under the Superfamily level are further grouped according to structural similarity in Structural Clusters (SCs). In recent years, CATH was expanded to classify domains under Superfamily level, according to functional similarity, i.e. grouping domains into Functional Families (FunFams) (Sillitoe et al., 2013).

The FunFam protocol involves profile-profile-based clustering of each domain sequence in each Superfamily to identify functional families (Sillitoe et al., 2013; Das et al., 2015a). The algorithm recognises evolutionary signals in cluster alignments, such as highly conserved positions and specificity-determining positions, which are used to improve functional coherence. For this, functional information is also used and obtained from resources such as GO (Harris et al., 2004), EC (Bairoch, 2000), and Pfam (Bateman et al., 2004). Hidden Markov Models (HMMs) profiles (Eddy, 1998) built from FunFams structural seed multiple sequence alignments (MSAs) were recently made available for function prediction and protein annotation in a new protocol called FunFHMMer (Das et al., 2015b). As a result of these developments, both structure and (predicted) sequence domains are now included in FunFams, which enables the analysis of domains and particularly domain interactions in the context of a larger protein universe. This helps to increase the CATH coverage of the current set of protein structures in the PDB, as well as in the prediction/classification of domains in newly solved structures or entire proteomes (which include many proteins with no structure available).

1.5.1.2 SCOP

In SCOP, the first level in the hierarchy is Class with five major classes: all α-helix, all β-strand, α/β, $\alpha + \beta$ and multi-domain (structures that are composed of two or more domains that belong to different classes). Next is Fold (there is no Architecture in SCOP). The Fold group describes how SSEs are arranged and their connectivity (topology). Proteins in the same SCOP fold share the same relative arrangement of major SSEs with the same topology. Although proteins in the same fold from different superfamilies may share a common evolutionary origin, such relationships typically cannot be distinguished from rare cases of convergent evolution. Then, the Superfamily level groups domain structures thought to be evolutionarily related. The Family level groups closely related domains which are likely to have similar structures and functions. Proteins in the same SCOP Superfamily are believed to share a common evolutionary origin despite having low or undetectable sequence similarity. SCOP's last biggest update was performed in June 2009, but the project is currently being developed as two distinct hierarchies, SCOP2 (Andreeva et al., 2014) and SCOPe (Fox et al., 2014).

1.5.2 Multiple domain definitions

Domain assignments vary according to the domain definition used. For example, Hadley and Jones, 1999, Veretnik et al., 2004, and Jefferson et al., 2007a, investigated the differences between CATH and SCOP structure-based domain assignments and found that there are instances where a domain in one classification has no equivalent in the other, despite the overall high degree of correspondence between

the two classifications. Domains within the same CATH or SCOP Superfamily share structural and functional similarities that suggest a common evolutionary ancestry but may not share a detectable sequence similarity. There are some Superfamilies that, despite having a significant number of sequence diverse relatives (<35% PID), have a high degree of structural conservation. A classic example is that of SUMO and Ubiquitin which show high structural similarity despite low sequence identity (Gill, 2004). Domain relatives within these Superfamilies have highly similar functions, and it is likely that the structural conservation is largely due to functional constraints.

1.5.3 Sequence-based domain predictions

CATH and SCOP classification hierarchies have been used to detect domain sequence relatives in genome and protein sequence databases. CATH predictions are provided by Gene3D (Lees et al., 2012). An average coverage of 50% of domain sequences is achieved for an average genome, which is even higher for bacterial organisms. For SCOP, SUPERFAMILY provides similar genome-wide predictions (Gough et al., 2001). Unlike CATH and SCOP which classify domains according to structure, Pfam (Bateman et al., 2004) classifies domains based on sequence similarity. For Pfam, domain predictions are also available for multiple genomes. There are ∼25% of domain sequences within a genome that can be assigned to structurally uncharacterised Pfam families.

1.6 Protein interactions

Most biological mechanisms, including transcription, translation, cell cycle control, signal transduction, transport, cellular motion, and secretion, are mediated by protein-protein interactions (PPI) and protein-ligand interactions (PLI) (Keskin et al., 2008; Pazos and Valencia, 2008; Khan et al., 2011). Therefore, determination of an organism's protein interactions (the interactome) is fundamental to decipher the mechanisms underlying cellular function (Okada et al., 2005). Additionally, many disease states are associated with protein interactions. A greater understanding of an organism's interactome network aids in determining which proteins are involved in disease, and has received much interest based on the potential for drug-targeting protein interactions (Blundell et al., 2006; Archakov et al., 2003).

Since the seminal work by Chothia and Janin, 1975, several studies have been carried out on the physicochemical characterisation of protein interfaces to understand the rules of molecular recognition. General principles of protein-protein interactions have been proposed (Hubbard and Argos, 1994; Jones and Thornton, 1996; Valdar and Thornton, 2001; Gao et al., 2004; Bahadur et al., 2004). Many authors defined measures for interface size, shape complementarity, protrusion, segmentation and secondary structure. Overall, interfaces have a tendency to be polar, uncharged and hydrophobic; and usually display a planar protruding shape, good shape complementarity and an overall high solvent accessible area (Jones and Thornton, 1996; Janin and Chothia, 1990; Jones and Thornton, 1997; Hubbard and Argos, 1994; Argos, 1988). Protein-protein interactions are held together by non-covalent forces including hydrogen bonds (between polar groups), ionic attractions (between charged

groups, also known as salt-bridges), Van der Waals forces (between molecules that have been polarised into dipoles), and hydrophobic interactions (among non-polar groups in aqueous solution).

Interacting proteins often have large surface patches that are in direct contact with each other. Especially enzymes, that select their substrates specifically, exhibit a high degree of shape complementarity onto their substrates. To account for conformational flexibility of protein-protein interactions, Koshland, 1958, proposed the 'induced fit theory' as a revision to the lock and key model first postulated in 1894 by Emil Fischer (Fischer, 1894). According to the latter theory, substrates are not rigidly docked to their enzymes, but constantly perform small rearrangements of their side chains. Shape complementarity is a purely structural feature and hence does not translate directly into sequence representations. However, shape complementarity can be the result of a long evolutionary process involving mutations on one side and compensatory mutations on the opposite side of the interaction. These correlated mutations are detectable at sequence level (Halperin et al., 2006; Marks et al., 2012; Pazos and Valencia, 2008).

1.6.1 Properties of protein interfaces

Protein interactions can be categorised as inter-molecular or intra-molecular, homodimers or heterodimers, transient or permanent and obligate or non-obligate (Jones and Thornton, 1996; Nooren and Thornton, 2003b; Ofran and Rost, 2003). A protein-protein (or domain-domain) interaction is classified as homodimer interaction if the interaction occurs between two identical chains or as a heterodimer interaction if the interaction is between two non-identical chains. A domain-domain

interaction can also be classified as a homodimer or heterodimer interaction depending on whether the two domains are identical or not. A domain-domain interaction is classified as intra-molecular if both domains are from the same polypeptide chain or intermolecular if the domains are from two different polypeptide chains. An interaction is classified as transient if its components associate and dissociate *in vivo*. In contrast, a permanent interaction is usually very stable and only exists in its complexed form. An interaction is classified as obligate if its components are not found as stable structures on their own *in vivo* whereas the components of a non-obligate interaction are found as stable structures. Structurally or functionally obligate interactions are usually permanent, whereas non-obligate interactions may be transient or permanent (Nooren and Thornton, 2003b). Unfortunately, it can be difficult to classify some interactions into these distinct types as proteins will interact in different ways depending on the cellular environment. Besides, transient interactions cannot be observed by any of the traditional structural determination methods.

Lo Conte et al., 1999, analysed protein-protein interaction sites from a range of different complex types including antigen-antibody complexes, protease-inhibitor complexes, complexes involved in signal transduction and enzyme-inhibitor complexes. Each of these complexes was formed by transient interactions where each of the proteins involved in the interaction was observed independently of the complex. On average protein-protein interaction interfaces were found to have the same non-polar nature as the rest of the surface of the proteins. The authors found that the size of the interface was related to the conformational changes that occurred upon binding. Large interaction interfaces were observed when the binding of the

complexes caused large conformational changes. Bridging water molecules are those forming two hydrogen bonds, one to each interface side, and participate extensively in hydrogen bonding at the interfaces (Teyra et al., 2011). Water molecules were found to contribute to the binding in two ways: affecting the close packing of atoms to ensure complementarity between the two surfaces and providing polar interactions between the two interacting surfaces (Lo Conte et al., 1999).

Surface patches with high hydrophobicity are energetically unfavourable in an aqueous solution, but favourable when in contact with other hydrophobic surfaces. Hence, their occurrence can be linked to binding sites. Gallet et al., 2000, proposed an interface detection method that predicts binding sites by analysing the hydrophobicity distribution in sequences. In particular, permanent interactions between globular proteins were found to involve more hydrophobic residues. The largest hydrophobic surface patches were often found to participate in protein binding, at least to some extent. Additionally, binding sites frequently show opposite charges in opposite interfacing patches.

In 2003, Nooren and Thornton, investigated the properties of transient protein-protein interaction interfaces (Nooren and Thornton, 2003a). They compared experimentally validated transient homodimers which are known to exist as monomers and dimers at physiological concentration to more stable, functionally validated, transient (i.e. intracellular signalling) heterodimers. The weak interactions had dissociation constants in the micromolar range whereas the most stable interactions had binding affinities in the nanomolar range. The weak interactions were found to be more planar and polar and to have a smaller contact size than the stronger interactions. The stronger interactions also showed large conformational changes

upon association and dissociation. According to Nooren and Thornton, 2003, hydrophobicity has less discriminative power for transient interactions, as transient interactions are often established by hydrogen bonds from polar side chains.

Ofran and Rost, 2003, used a much larger data set to analyse different domain-domain interfaces. They identified six different types of interfaces and found that using just residue type or pairwise residue type they could discriminate between different types of interface. The six interface types were within the same structural domain, between different domains, permanent, transient, homo-oligomers and hetero-oligomers. All of the interfaces differed significantly from background residue compositions, surface residues and internal residues. They predicted which of the six types of interfaces a pool of 1000 residues belonged to with an accuracy ranging from 63 to 100%.

Zhanhua et al., 2005, compared the interface properties of homodimers and heterodimers. The authors found that homodimers typically have a greater number of interface residues and hydrogen bonds but the density of hydrogen bonds was greater for heterodimers. They also found that charged hydrophilic residues were dominant at heterodimer interfaces, in contrast to a dominance of hydrophobic residues at homodimer interfaces.

Calculation of the electrostatic potential of protein-protein complexes revealed that protein-protein interfaces display electrostatic complementarity (McCoy et al., 1997). With the exception of covalent bonds, the bonds formed by hydrogen atoms between a hydrogen donor and a hydrogen acceptor are the strongest contributors to binding energies (McDonald and Thornton, 1994; Panigrahi and Desiraju, 2007).

1.6.2 Key binding residues at the interfaces

Hot spots are interface residues that dominantly contribute to the binding free energy. Their identification requires experimental techniques such as alanine scanning (Bogan and Thorn, 1998). Ma et al., 2003, found that structurally conserved residues distinguish between binding sites and exposed protein surfaces. Halperin et al., 2004, showed that hot spots are often observed to couple across protein-protein interfaces. According to Keskin et al., 2005, hot spots reside in tightly packed, structurally conserved regions that contribute dominantly to the stability of the interaction.

Structurally and functionally important residues are often well conserved (Hu et al., 2000; Guharoy and Chakrabarti, 2010; Fiser et al., 1996). The identification of binding sites based on evolutionary conservation, however, remains controversial. While some authors claim a stronger conservation of interfaces than the rest of the surface (Nooren and Thornton, 2003b; Bordner and Abagyan, 2005), others question the statistical significance (Caffrey et al., 2004). A remaining problem is how to clearly identify non-functional parts of the protein surface, as the knowledge about interactions might be incomplete. Surface parts considered non-functional could be important for yet unknown interactions. It is hence not sufficient to base binding site detection solely on conservation scores. However, selective pressure often leads to the conservation of important protein features, such as binding behaviour. Therefore, it is generally beneficial to consider evolutionary conservation scores (Li et al., 2004; Bordner and Abagyan, 2005).

1.6.3 Experimental methods

Experimental methods to identify protein interactions can be classified into two broad areas: traditional experimental methods and high-throughput methods. Low throughput experimental techniques are highly accurate but time-consuming. The high-throughput methods can determine genome-wide protein-protein interactions in a fraction of the time taken by the traditional methods but are traditionally less reliable. Among the traditional experimental methods are: immunoprecipitation (Berggård et al., 2007a; Sambrook and Russell, 2006); FRET (Fluorescence Resonance Energy Transfer) (Kenworthy, 2001); Mass Spectrometry and Quantitative Proteomics (Berggård et al., 2007b; Kirkwood et al., 2013); and Tandem Affinity Purification (TAP)-tags (Berggård et al., 2007b). Among the high-throughput approaches is the yeast two-hybrid (Y2H) method (Fields, 2005). Although originally developed as a low-throughput method, Y2H has been developed into a high-throughput method for determining binary protein-protein interactions. The Y2H method has been applied to the human genome on a wide scale (Stelzl et al., 2005). As many as 90% of known protein-protein interactions are thought to be missed by the genome-wide Y2H projects (Ito et al., 2001) and both high false positive rate and false negative rate also been observed (Sprinzak et al., 2003; von Mering et al., 2002; Deane et al., 2002).

1.6.4 Computational methods

The field of protein-protein interaction prediction is one area of functional annotation of proteins which has received a significant amount of attention. A wide range

of computational methods has been developed. Many of these methods are designed to improve the reliability of experimental methods, but others directly predict new protein-protein interactions. The methods which directly predict new interactions can be classified into genome-based, sequence-based and structure-based methods, or a combination of them.

A field of protein-protein interaction prediction which particularly depends on structural data is molecular docking. Molecular docking aims to predict *ab initio* the structure of complexes from their component domains when the structure of the individual component domains is known. Most molecular docking methods assume that the proteins interact and try to determine the complex structure of the two proteins when they are bound. There are many difficulties in calculating the protein-protein interaction energetics directly from the coordinates of the complex. A significant challenge is to cope with protein flexibility (Bonvin, 2006). Although obtaining a precise atomic-level model of a docked pair of proteins is a difficult problem, the results of CAPRI docking prediction experiments suggest progress has been made for complexes where few conformational changes occur on binding (Janin et al., 2003).

1.6.5 Databases of protein interactions

A more detailed picture of protein interactions is available from the analysis of protein structures. This has prompted the use of 3D protein complexes to predict interactions for homologous proteins. In addition to the basic information that two proteins interact, 3D structures also provide detailed information about the specific residue (or atoms depending on the resolution) contacts that mediate the interaction

of two proteins. This is a significant advantage over other protein-protein interaction predictive methods resulting in a more detailed data source that can be employed to analyse the properties of protein-protein interfaces and analysis of variation.

The analysis and prediction of protein interaction sites from structural data is limited by the availability of structural complexes. Several methods and accompanying databases that use structure data to inform and predict protein interaction interfaces have been developed. Among the most prominent are: 3did (3D interacting domains) (Mosca et al., 2014), PIBASE (Davis and Sali, 2005), SCOPPI (Winter, 2006), PSIBase (Gong et al., 2005), PRISM (Baspinar et al., 2014), ModBase (Pieper et al., 2011) and iPfam (Finn et al., 2014b). These methods rely heavily on structural matching and homology. The principle is that if two proteins A and B are seen to interact in a structure complex, then a homologue of A is predicted to interact with a homologue of B.

A common thread among the newest databases, such as DIP (Salwinski et al., 2004) and STRING (Franceschini et al., 2013), is the combination of prediction methods and data from multiple sources, as well as literature mining. DIP employs such methods to annotate protein-protein interaction data in addition to manual review from expert biologists before submission to the database. STRING predicts interactions based upon high-throughput experimental data, from the mining of databases and literature, and from predictions based on genomic context analysis. STRING annotates each prediction with a degree of certainty. The certainty of interaction increases with the number of predictive methods which suggest interaction.

1.7 Genetic Variation

Recent advances in DNA sequencing technology, the so-called next-generation sequencing (NGS), indicate that it is increasingly routine to carry out whole-genome sequencing on multiple individuals (Choi et al., 2009; Soon et al., 2013). NGS projects have generated a very large catalogue of human variants, but the interpretation of these data remains challenging. It is difficult to determine the functional impact of genetic variation on individuals (Capriotti et al., 2012; Mirnezami et al., 2012) and populations (Schofield and Hancock, 2012; Blair et al., 2013). These difficulties have become more pronounced and important as the scope of analysis has expanded from monogenic disorders (Chong et al., 2015; Ng et al., 2009) to complex diseases (Ward and Kellis, 2012). Improvements in sequencing technologies have also allowed moving beyond studying associations in the exome. Several large-scale genome projects have provided evidence that at least 80% of the human genome is functional (ENCODE Project Consortium, 2012).

One prominent result of massive sequencing projects such as the 1,000 genomes project (1kGP) (The 1000 Genomes Project Consortium, 2012) is an ever increasing diversity and number of genetic variants that are identified in both protein and non-protein coding regions of the genome. Single nucleotide polymorphisms (SNPs) are among the most common types of genetic variation, accounting for 90% of known sequence differences. Although many of these changes are neutral, currently there are more than 100 million SNPs, the majority being identified for human, which give rise to a large number of amino acid substitutions in proteins (The 1000 Genomes Project Consortium, 2012). One of the major SNP databases is dbSNP (Sherry

et al., 2001). Figure 1.5 shows the exponential increase in the number of both Reference SNP clusters and validated SNPs, deposited in NCBI dbSNP database since 2002.

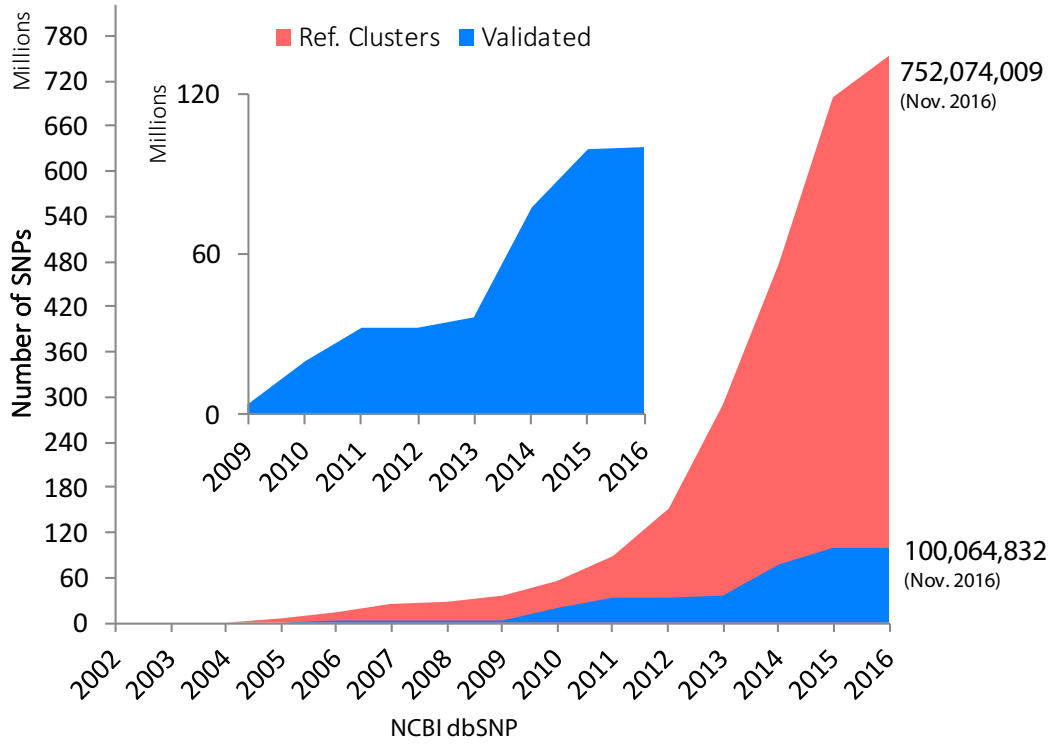


Figure 1.5: The exponential growth in the availability of genetic variation data in dbSNP. dbSNP is provided by the NCBI and is one of the oldest and biggest repositories for genetic variants. dbSNP contains validated entries and an exponentially increasing number of reference SNP clusters. Data obtained from `http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi` as accessed in November 2016.

1.7.1 Types of genetic variation

Genetic variation occurs both within and among populations, supported by individual carriers of the variant genes. It can be divided into different forms according to the size and type of genomic variation underpinning genetic change (Ward and

Kellis, 2012; Zhang et al., 2010; Stankiewicz and Lupski, 2010; Huminiecki and Conant, 2012; Kamburov et al., 2015). These include: small-scale sequence variation (<1Kbp) such as substitutions and indels (Zhang et al., 2010); as well as large-scale structural variation (SV) (>1Kbp) (Stankiewicz and Lupski, 2010; Gonzaga-Jauregui et al., 2012), such as copy number variation (CNV) (copy number loss and copy number gain) and rearrangement (translocation and inversion). Additionally, there are other massive gene variation events that lead to changes in the number of chromosomes (Huminiecki and Conant, 2012) (e.g. whole-genome duplication), which can result in polyploidy (Huminiecki and Conant, 2012) or aneuploidy (Kamburov et al., 2015).

Variation can be categorised according to its location in the genome in coding regions of genes, non-coding regions of genes or in the intergenic regions. The majority of SNPs occur in non-coding regions (Syvänen, 2001). However, SNPs that do occur in the coding region are important because they can affect a variety of important biological and molecular activities such as stability (Casadio et al., 2011), expression level (Heinzen et al., 2012) and protein function (Chasman and Adams, 2001). Nonetheless, recent studies (Prado-Montes de Oca et al., 2009; Visel et al., 2010; Ward and Kellis, 2012) have shown that SNPs in non-coding regions may still affect other activities such as gene splicing, gene expression and the sequence of non-coding RNA.

It has been suggested that 20% of non-synonymous SNPs (nsSNPs) could potentially damage proteins (Sunyaev et al., 2001). SNPs affect how an individual responds to diseases (Wellcome Trust Case Control Consortium, 2007), drugs (Giacomini et al., 2007; Lahti et al., 2012; Ma and Lu, 2011) and environmental factors

(Bresciani et al., 2013). On average, human DNA consists of a SNP for every 300 bases (Nelson et al., 2004). This means that for the whole genome (∼3 billion bases) there would be roughly 10 million SNPs. More than 60,000 (35%) SNPs are located in the coding regions of the genes (Sachidanandam et al., 2001). Half of these (nsSNPs) cause amino acid substitutions (Cargill et al., 1999). Contrastingly, an nsSNP alters the amino acid sequence of a protein, while a synonymous SNP is a SNP that does not change the sequence. Synonymous SNPs can, however, affect the expression of the gene product by interfering with normal mRNA splicing, leading to abnormally short or long gene products. Synonymous coding SNPs may also cause alterations in mRNA folding and translation of the protein (Kudla et al., 2009).

An nsSNP can be either missense or nonsense. A missense nsSNP results in a different amino acid, while a nonsense nsSNP change results in a premature stop codon. Additionally, a missense nsSNP can either be a conservative or non-conservative change. A non-conservative change results in a different residue with substantially different physicochemical properties. Figure 1.6 illustrates the process of protein translation and the main common variation events that take place throughout the genome. Variation types included are: 1) regulatory; 2) splicing; 3) missense; 4) synonymous SNP; and 5) nonsense variants.

1.7.2 Experimental methods

Different methods have been developed to detect SNPs. SNP detection has been traditionally performed by sequence scanning and genotyping. Experimental techniques include: Denaturing High-Performance Liquid Chromatography (DHLPC) (Yu et al., 2001), direct DNA Sequencing (Berg et al., 1995), Microarray genotyping

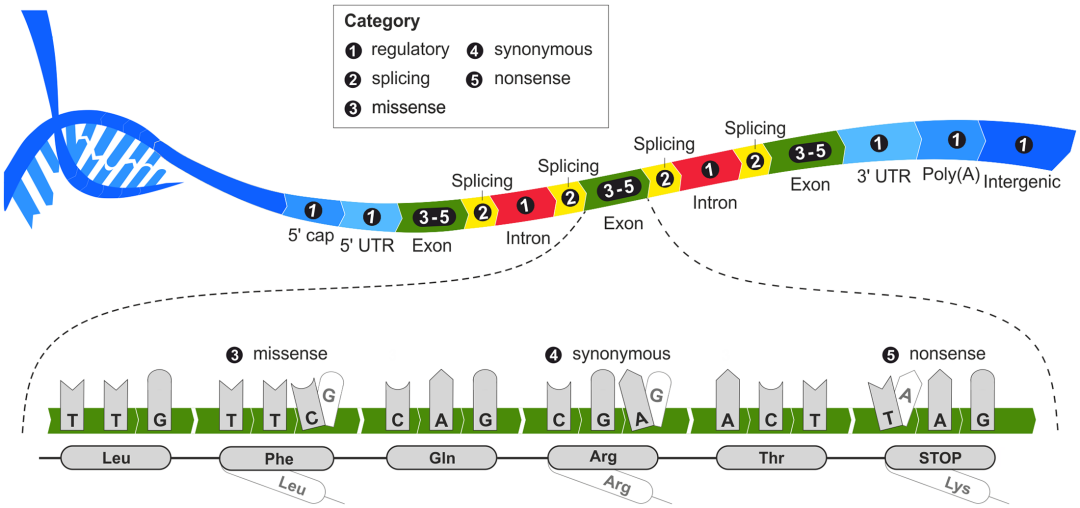


Figure 1.6: Different types of genetic variants according to their location in the genome. Variants are categorised as: 1) regulatory; 2) splicing; 3) synonymous; 4) missense; and 5) nonsense. Figure from Bendl et al., 2016.

(Maskos and Southern, 1992), and Mass Spectrometry (MS) (Ross et al., 1998). Recent approaches to detecting genetic variants resulted from the massive improvements in NGS technologies. Variant calling statistical methods are now routinely applied to large-scale NGS sequencing data (Kwok and Duan, 2003; Nielsen et al., 2011; Cheng et al., 2014). In addition to methods which detect germline genetic variants by aligning reads from individual samples to a reference genome, somatic variants can also be detected from multiple tissue samples within a single individual. These variants correspond to mutations that are not present within the individual's germline cells but occurred *de novo* within groups of somatic cells. Many studies on cancer are designed around investigating the profile of somatic mutations within cancerous tissues. However, because of the high-throughput nature of most of these efforts, many polymorphisms have not been experimentally characterised in terms of their possible disease association.

1.7.3 Main repositories for genetic variation

The most widely used variation databases are: dbSNP (Sherry et al., 2001), 1kGP (The 1000 Genomes Project Consortium, 2012), HapMap (The International HapMap Consortium, 2007), HGMD (Stenson et al., 2014), COSMIC (Forbes et al., 2015), Phencode (Giardine et al., 2007), ExAC (Exome Aggregation Consortium (ExAC), Cambridge, MA, 2016), and ESP (NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, 2016). The majority of these databases include links to genome browsers and protein databases, to integrate and link genomic data and information about the gene product to clinical and phenotypic information. The majority of the known SNPs' data are stored in the dbSNP database hosted by the NCBI. The goal of dbSNP is to act as a primary database containing all known genetic variations, including the location in the chromosome, the allele frequency, the associated literature. Ensembl (Chen et al., 2010; Flicek et al., 2013; Cunningham et al., 2015) and UniProt (The UniProt Consortium, 2015) are currently collecting and organising genetic variation from multiple databases (Chen et al., 2010; The UniProt Consortium, 2015), which integrate well with the sequence information databases and functional annotations they provide. This makes them the go-to providers for obtaining comprehensive sets of annotated variants.

Among the most notable projects and variation resources are the 1kGP and the UniProt Humsavar dataset. The 1kGP (The 1000 Genomes Project Consortium, 2012) is an international research collaboration focusing on genetic variation in humans. The primary goal of 1kGP is to create a comprehensive catalogue of human genetic variation. Additionally, the aim is to estimate the population frequencies,

haplotypes and linkage disequilibrium patterns. 1kGP also aims to support better SNP calling and probe selection for genotyping platforms in future studies and for the improvement of the human reference sequence.

UniProt Humsavar (Wu et al., 2006; Famiglietti et al., 2014) is a manually curated list of SNPs in human proteins developed by the UniProt Consortium (The UniProt Consortium, 2015). The information is obtained from the literature, and the SNPs have been mapped to protein sequences. The Humsavar list consists of information such as the amino acid changes, the variant location in the protein sequence, the associated protein, the dbSNP identifier (if available) and the variant disease status, i.e. whether it is associated with disease or not. All of the SNPs in the Humsavar list are nsSNPs. The disease variant status is based on literature reports of disease association from the OMIM (Amberger et al., 2015). The OMIM catalogues known Mendelian inherited diseases known to occur in humans. More recently, the ClinVar database (Landrum et al., 2016) was developed to provide relationships among medically important variants and phenotypes. It contains a more exhaustive list of curated variants when compared to UniProt Humsavar. Other specialised databases such as SAAPdb (Hurst et al., 2009), contain lists of structurally-mapping variants from multiple sources.

1.7.4 Consequences of genetic variation

There is a whole range of effects which genetic variation can have on the phenotype of an individual. Mutations are usually classified into three groups, based on the fitness change: 1) beneficial mutations increase the fitness of an individual; 2) neutral ones lack a visible effect on the fitness; and 3) deleterious or pathogenic mutations that

lower the overall fitness. Missense mutations account for approximately half of all allelic variants underlying Mendelian human diseases (Stenson et al., 2014; Krawczak et al., 2000; Hamosh, 2004). Thousands of rare heritable Mendelian disorders have been linked to genes in which a single amino acid variation is both necessary and sufficient to cause disease (Hamosh, 2004). In contrast, the combined effect of different susceptibility genetic variants and environmental factors are thought to result in common multi-factor diseases which have proven much more challenging to study. Figure 1.7 shows a selection of disease-causing mutations that affect a variety of proteins including uroporphyrinogen decarboxylase (Figure 1.7 A), von Hippel-Lindau protein (Figure 1.7 B and C), p53 (Figure 1.7 D), von Willebrand factor A1 (Figure 1.7 E), Factor IX (Figure 1.7 F and G), and transthyretin (Figure 1.7 H-I).

A missense mutation located at a site critical to protein function typically leads to a disease phenotype. A critical site may be a catalytic residue or a residue involved in ligand binding in an enzyme, or a residue involved in binding to partner molecules. The disease phenotype in these cases may arise because of loss or gain of function, or altered protein binding specificity or affinity, while the expression or stability of the protein product is not necessarily affected. Translation initiation codons can be affected by missense mutations as well, preventing the formation of the protein product, or translation may start at the next possible ribosome starting point, in which case the protein product would be truncated. The consequences of missense mutations affecting functional sites are rather straightforward to define when the protein in question is well known because information regarding the critical residues is typically annotated in major protein databases. The Catalytic Site Atlas

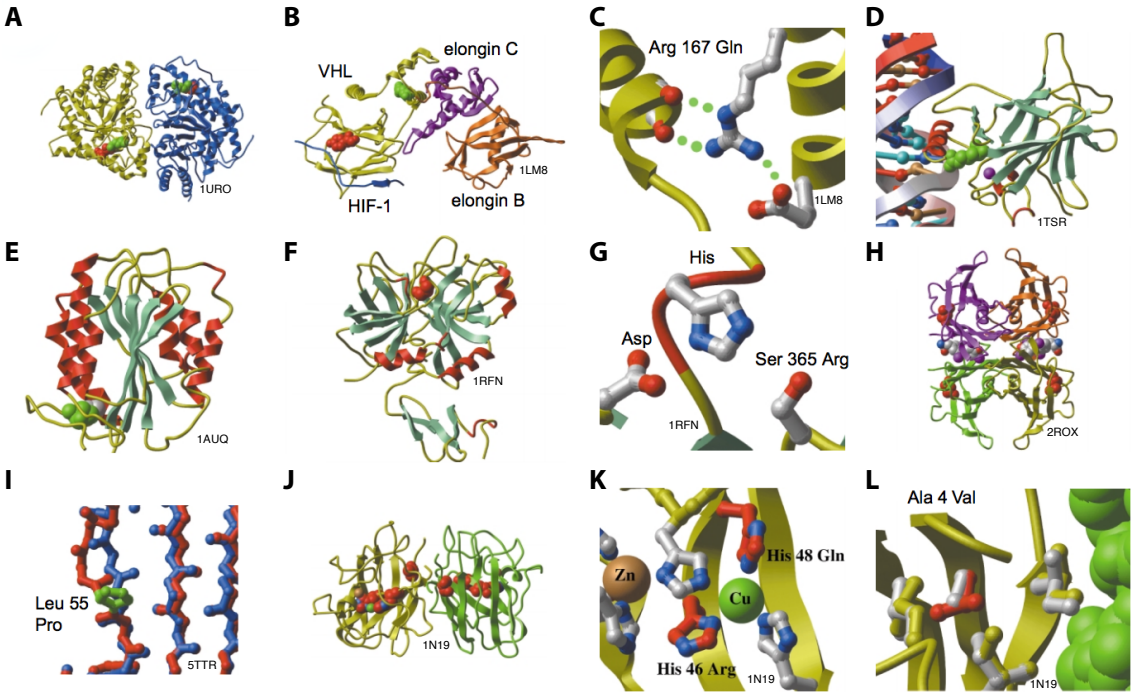


Figure 1.7: Examples of pathogenic mutations mapped onto protein structures. A) Mutation of uroporphyrinogen decarboxylase homodimer Met165Arg (green) and Leu195Phe (red). B) and C) Mutation of the von Hippel-Lindau protein Tyr98His (red) and Arg167Gln (green) in complex with elongin B, elongin C and a HIF-1 peptide. D) The p53 tumour suppressor DNA-binding domain Arg273His (green) in complex with DNA. E) The von Willebrand factor A1 domain Cys509Arg (green). F and G) Mutation of Factor IX Ser365Arg (red). H and I) The transthyretin tetramer Leu55Pro (red and blue) and in complex with two molecules of thyroxine bound between subunits. J-L) Cu-Zn superoxide dismutase His46Arg, His48Gln, Ala4Val in complex with zinc (orange) and copper (green) ions. For more details refer to Steward et al., 2003. Figure adapted from Steward et al., 2003.

(Porter et al., 2004) is a specialised database for detailed annotations of known and predicted enzyme catalytic residues.

Disease-associated mutations change the function or the structural stability proteins, whereas residue differences between evolutionarily related proteins usually conserve protein structure and function (Steward et al., 2003; Vitkup et al., 2003). Pathogenic mutations tend to occur at positions conserved between species in evolution (Steward et al., 2003; Sunyaev et al., 2001; Shen et al., 2006; Miller and Kumar,

2001; Ferrer-Costa et al., 2005), and classically, highly conserved positions in multiple sequence alignments often point to functional sites (Capra and Singh, 2007; Panchenko et al., 2004; Chung et al., 2006). When considering structural mutations, the level of conservation of the physicochemical properties between the wild type and the substituting amino acid has an effect on the pathogenicity of the mutation, so that conservative substitutions tend to be less frequently pathogenic than those significantly altering the residue properties such as charge, hydrophobicity, or size (Miller and Kumar, 2001; Tang et al., 2004; Stone and Sidow, 2005; Khan and Vihinen, 2007; Briscoe et al., 2004). The hydrophobic nature of residues located in the core of a protein tends to be conserved, and these residues can usually be identified in multiple sequence alignments.

On the contrary, there is a low abundance of disease-causing mutations occurring at positions that change in evolution (Miller and Kumar, 2001; Briscoe et al., 2004). Consequently, sequence conservation and phylogenetic studies are powerful for the prediction of functionally and structurally important residues in proteins. However, mouse genome data revealed many pathogenic mutations in humans that are wild-type residues in mouse orthologues (Mouse Genome Sequencing Consortium, 2002), so it cannot be directly assumed that because the same residue type appears at equivalent positions in close homologues, it will not lead to disease in humans. This observation is partially explained by the fact that co-evolution of the sites vital to the structure and/or function of a protein is rather common. When a critical site is mutated, a compensating mutation occurs at a site that is functionally, energetically of physically linked to that position. Analysis of covariant positions in multiple sequence alignments may thus reveal conserved positions that are relevant to the

function or structure of a protein (Halperin et al., 2006; Marks et al., 2012; Pazos and Valencia, 2008).

Buried positions are more sensitive to pathogenic mutations than positions on the surface of the protein, because alterations in such positions, especially in the hydrophobic core of a protein, have the potential to cause greater disruption of the overall structure of a protein (Terp et al., 2002; Steward et al., 2003; Sunyaev et al., 2001; Vitkup et al., 2003). Residues at these positions form critical stability-maintaining contacts with other residues, and these may be disrupted when a residue is altered. Changes in the size (Buckle et al., 1996; Liu et al., 2000; Loladze et al., 2002; Matthews, 1993), hydrophobicity (Liu et al., 2000; Matthews, 1993; Shortle et al., 1990), or charge of the residue side chain at buried positions usually have an effect on the structural stability of the protein (Chasman and Adams, 2001; Sunyaev et al., 2001). Mutations at solvent accessible sites might interfere with the interactions a protein forms with other molecules, or they may contribute to the solubility or stability of a protein (Gribenko et al., 2009; Sunyaev et al., 2001; Wang and Moult, 2003; Strickler et al., 2006).

1.7.5 Predicting the consequences of genetic variation

Understanding the molecular consequences of the mutations that cause human genetic disease remains an important research challenge (Steward et al., 2003; Mooney, 2005; Ng and Henikoff, 2006; Karchin, 2009). Over the last years, several research groups have developed computational techniques to predict the deleterious effects of missense mutations (Dobson et al., 2006), as it aids for prioritising experimental variant analysis. Methods range from those that use sequence information together

with known features such as functionally important residues, secondary structure and the location of buried residues (for example SIFT (Ng and Henikoff, 2003)), to those that consider protein 3D structures (for example Polyphen-2 (Adzhubei et al., 2013), SDM (Worth et al., 2011), I-Mutant 2.0 (Capriotti et al., 2005) and SNPs3D (Yue et al., 2006)). Several methods have been developed by integrating the results of previously developed prediction tools (for example Condel (González-Pérez and López-Bigas, 2011) and PredictSNP (Bendl et al., 2014)). A couple of methods were also developed specifically to predict changes in protein-protein binding (BeAtMuSiC (Dehouck et al., 2013) and SNP-IN (Zhao et al., 2014)). In addition to prediction methods, some effort has been put into the functional analysis of genetic variants. These range from evolutionary models for examining the mechanisms for generating and transmitting variation in evolving populations (Rivoire and Leibler, 2014), to studies of the relevance of synonymous mutations as a driver in human cancers (Supek et al., 2014).

SIFT and Polyphen-2 are the two most commonly used variation classifiers. SIFT (Ng and Henikoff, 2003) uses sequence features to predict the effects of nsSNPs on proteins. The program is based on the observation that conserved residues tend to be more intolerant towards substitution than non-conserved residues. It estimates positions that will be unfavourable to mutation based on tolerated mutations in homologs. The SIFT program generates a list of homologous sequences using the PSI-BLAST algorithm (Altschul et al., 1997) before aligning them to score the probability for all the possible amino acid substitutions. Polyphen-2 (Adzhubei et al., 2013) uses a hybrid approach that considers both sequence and structure to distinguish between deleterious and non-deleterious nsSNPs. Polyphen-2 has an

additional three structure-based predictive features that it uses in its scoring system, the dihedral angles, secondary structure and accessible surface area (ASA).

1.8 Overview of this Thesis

The main aim of the work presented in this Thesis is to improve the current methods for the analysis of genetic variation in the context of rich structural information, focusing particularly at interaction interfaces. This aim was prompted by the growing availability of both structural and genetic variation data, which presented as a unique opportunity to push forward the scope of genetic variation analysis at protein structure domain interfaces. Additionally, despite the fact that interaction interfaces have been extensively studied, both computationally and experimentally, and the fact that the involvement of binding impairment resulting from variation has been linked to disease, there is tremendous scope for studying the structural and functional consequences of variation using such an unprecedented structural and variation coverage. This aim has been approached by developing integrative methods that allow structural and sequence data to be combined and enable the analysis of variation on domain interaction interfaces, within a structural and evolutionary perspective.

Chapter 2 explores the development of a new computational framework that allows structural and sequence data to be combined to enable the study of genetic variants on protein interfaces. ProIntVar (Protein Interactions and Variants) encompasses all the tools and methods necessary for the analysis of genetic variants in the context of feature-rich protein structural data. Chapter 2 also overviews the

datasets collected and organised in ProIntVar. ProIntVar provides a framework for the seamless analysis of sparse structural data, sequence data, interaction data and genetic variation data. It implements routines for generating biological assemblies, defining protein interactions, annotating additional structural features in sites and regions. It allows protein structure, protein sequence and genomic DNA sequence to be cross-mapped. It incorporates generated structure-based MSAs for CATH structural clusters and functional families, which are further extended with similar protein sequences (Chapter 3). Finally, it annotates and organises genetic variation from a large number of sources and repositories.

Considering domain-domain and domain-ligand interactions is a more reliable approach to perform analysis of the effects of genetic variation in protein complexes (Jones et al., 2000; Stein, 2004). To perform an enriched analysis of genetic variants under a structural and evolutionary perspective, structure-based MSAs are generated for CATH structural and functional domain families. Chapter 3 focuses on improving the quality of the structure-based MSAs by: 1) developing methods that take advantage of the features of STAMP (Russell and Barton, 1992); and 2) exploring the power of HMMs for extending the alignments and annotating them with homologous protein sequences. Hit sequences are believed to be structurally/functionally homologous and thus a rich source of novel insight into the structural context and potential consequences of a vast number of genetic variants.

Chapter 4 explores the overall analysis of genetic variants in protein families, in the context of various structural regions and environments. The in-depth characterisation of the variants across different environments by type of substitutions and annotation is important, as it enables finding transitions that might be implicated

in disease, and that can potentially affect protein stability, activity and function. Analysis of the genetic variation that maps onto domain-domain and domain-ligand interaction interfaces is further explored in Chapter 5.

Chapter 5 focuses on the overall analysis of genetic variants at structurally conserved interface positions, and on the classification of domain-domain interactions. The global trends of variation identified can be used to help prioritise variants and protein families for further analysis. Three hand-picked domain families showing variation mapped onto interaction sites were further analysed as proof-of-concept examples.

Chapter 2

Development of ProIntVar

2.1 Summary

This chapter overviews the general methods and the implementation details underlying ProIntVar (Protein Interactions and Variants), the main computational framework developed in this Ph.D. project. Since the development of ProIntVar and its contents are central to all the studies performed and reported in the following Chapters, ProIntVar is described in detail here. A brief comparison to similar databases/systems is performed, followed by a final overview of the main unique features of ProIntVar. Wherever necessary, the methods and datasets introduced in this Chapter are revisited in more detail in the next Chapters.

2.2 Introduction

The analysis and prediction of the consequences of protein mutation have received extensive attention from the scientific community in recent decades. Initial analysis

of protein mutations was performed in order to identify those associated with disease states. Various Mendelian traits were linked to single point mutations and the disease mechanisms identified (Steward et al., 2003; Krawczak et al., 2000; Hamosh, 2004; Stenson et al., 2014). The structural analysis of genetic variation was performed on a case by case basis, until recent reports that focused on the characterisation of these in hundreds of protein structures spanning dozens of protein families (de Beer et al., 2013; Porta-Pardo et al., 2015; Peterson et al., 2010; Vázquez et al., 2015). ProIntVar was developed in this work with the aim of characterising vast genetic variation datasets in the context of an unprecedented number of protein structures currently available. The general overview of the field necessary to understand the contents of this chapter was provided in the main introduction, Chapter 1. Thus, the next Section will go straight to the contents and methods implemented in ProIntVar.

2.3 Methods and Contents

2.3.1 Overview of ProIntVar

ProIntVar (Protein Interactions and Variants) has been developed in this work to allow structure-centric analysis of genetic variants in proteins. Figure 2.1 overviews the various components, i.e. methods and datasets, integrated in ProIntVar. ProIntVar enables integrative multiple-level mapping between proteins and their annotated domains to protein sequences and known nsSNPs, disease-associated variants and somatic mutations (Figure 2.1 A and D). The raw structural datasets obtained from the PDBe were cross-mapped to UniProtKB and Ensembl (via SIFTS (Velankar

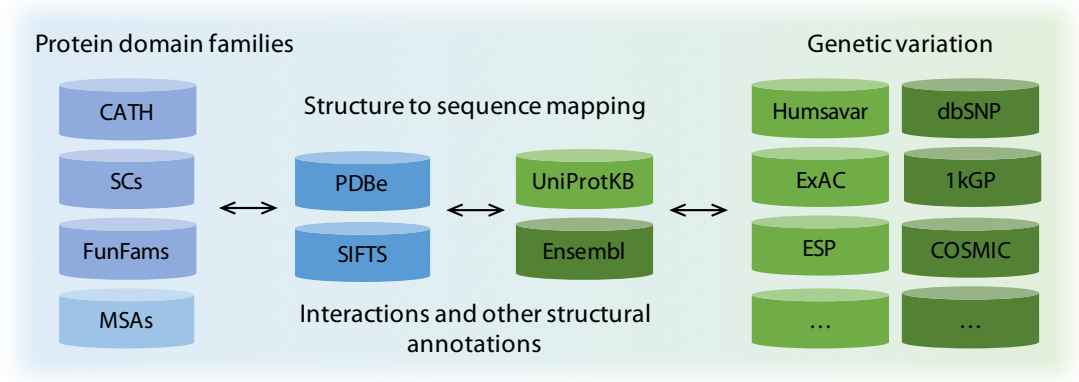
et al., 2013)) and further annotated with information on biological assemblies, interaction interfaces, secondary structure and relative solvent accessibility (RSA), as described throughout this Chapter (Figure 2.1 B). Domain annotations were provided from the CATH (Class, Architecture, Topology and Homologous Superfamily) resource (Sillitoe et al., 2015) for a subset of the protein structures. Additionally, ProIntVar includes structure-based multiple sequence alignments (MSAs) generated for protein domains annotated in CATH at structural clusters (SCs) and functional families (FunFams) level, under Superfamily (SPF) (Figure 2.1 C). The following Sections describe the full contents of ProIntVar in greater detail. In order to improve the understanding of how the methods and datasets were integrated in this work, Figure 2.2 shows a flowchart diagram highlighting the main raw data, datasets generated and processing steps performed by ProIntVar.

2.3.2 Collecting and organising structural data

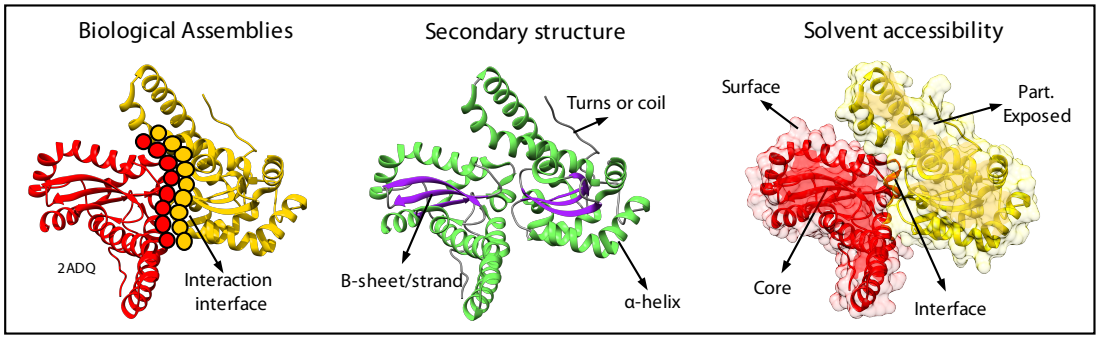
As described in Section 1.4, protein structures are deposited in the worldwide Protein Data Bank (wwPDB) central archive of macromolecular structures (Westbrook et al., 2002; Berman et al., 2003, 2007). To be able to work with the volume of structural data currently available, protein structures were downloaded from the PDBe's (Protein Data Bank in Europe) FTP server with the Rsync utility. Since the *PDB* file format was due to be phased out and replaced by mmCIF format in 2016, mmCIF was used as the standard structure format avoiding the disruption of making the transition later in the course of the project.

The various advantages of using mmCIF format are related to the way in which the file format stores and annotates the structural data. Although mmCIF was

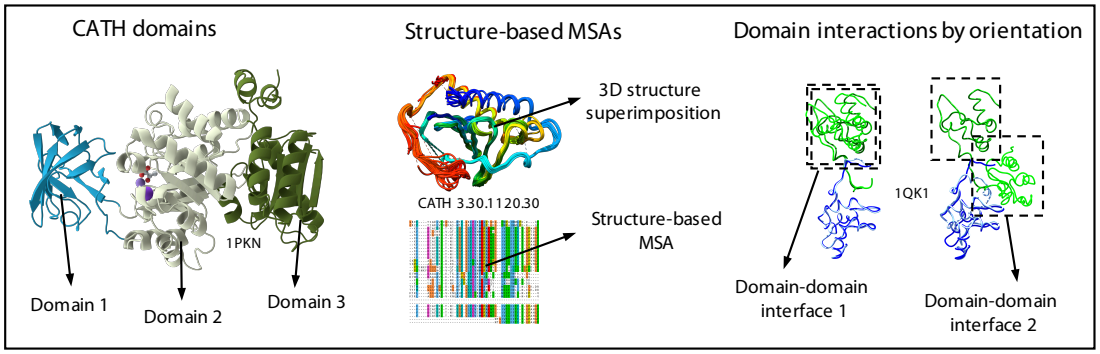
A Contents of ProIntVar (Protein Interactions and Variants)



B Protein structures, biological assemblies, interaction interfaces and structural annotations



C CATH protein families, domain definitions, structure-based MSAs and domain interactions



D Genetic variation mapped onto protein domain structures and interaction interfaces

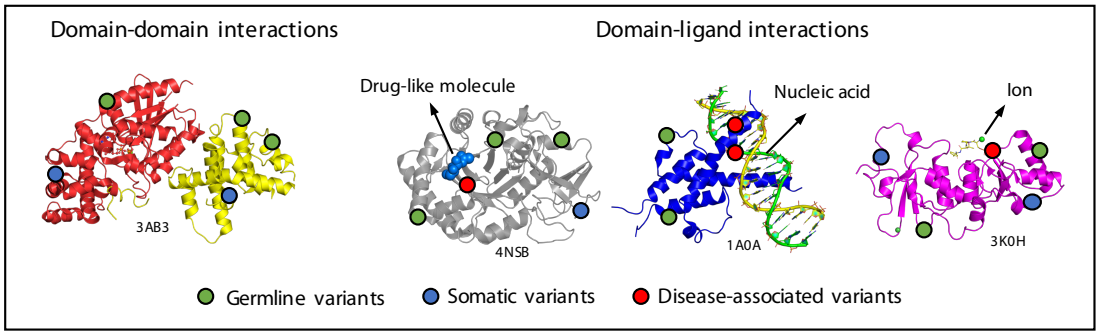
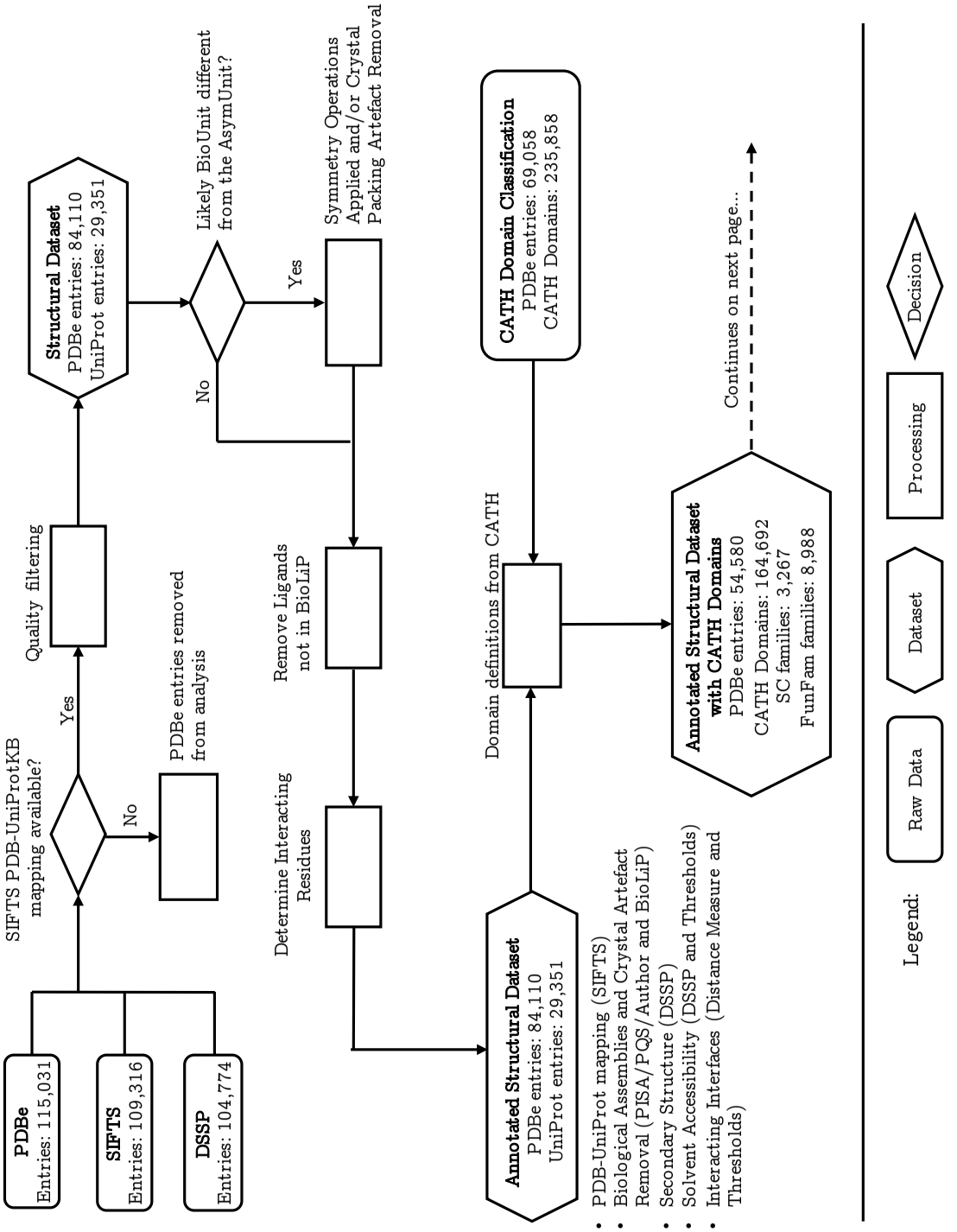


Figure 2.1: Overview of the main methods and resources integrated in ProIntVar. A) Protein structure to protein sequence mapping is performed with the aid of SIFTS. Protein sequence to genomic DNA coordinates is performed with the aid of Ensembl/UniProtKB cross-reference tools to allow mapping of genetic variants. B) ProIntVar integrates structural data from PDBe and secondary structure from DSSP, as well as additional structural annotations generated in this work. C) Domain definitions are obtained from CATH. Structure-based multiple sequence alignments (MSAs) are generated in ProIntVar with an optimised protocol that relies on STAMP. D) Missense variants are collected from several resources and databases described in Section 2.3.13, through the Ensembl REST API and the EBI UniProt Variation API. See the text for more details.

originally developed to describe small molecule structures, the format has been extensively developed to represent and describe larger macromolecular structures, such as protein complexes and nucleic acids. In fact, mmCIF is ideal to represent large structures which cannot be fully represented in the PDB file format, due to historical limitations. Large structures previously split into multiple PDB files, are provided in a single mmCIF file from 2016 onwards. The mmCIF format accommodates table-like data structures and key-value data entries all defined by a descriptive set of dictionaries for data categories, category groups and data items. For each of these entities, the SMCRA (Structure, Model, Chain, Residue, Atom) model still applies (Hamelryck and Manderick, 2003).


PDBe
Entries: 115,031
SIFTS
Entries: 109,316
DSSP
Entries: 104,774
SIFTS PDB-UniProtKB mapping available?
Yes
No
PDBe entries removed from analysis
Quality filtering
Structural Dataset
PDBe entries: 84,110
UniProt entries: 29,351
Likely BioUnit different from the AsymUnit?
Yes
No
Symmetry Operations Applied and/or Crystal Packing Artefact Removal
Remove Ligands not in BioLiP
Determine Interacting Residues
Annotated Structural Dataset
PDBe entries: 84,110
UniProt entries: 29,351
• PDB-UniProt mapping (SIFTS)
• Biological Assemblies and Crystal Artefact Removal (PISA/PQS/Author and BioLiP)
• Secondary Structure (DSSP)
• Solvent Accessibility (DSSP and Thresholds)
• Interacting Interfaces (Distance Measure and Thresholds)
CATH Domain Classification
PDBe entries: 69,058
CATH Domains: 235,858
Domain definitions from CATH
Annotated Structural Dataset with CATH Domains
PDBe entries: 54,580
CATH Domains: 164,692
SC families: 3,267
FunFam families: 8,988
Continues on next page...
Legend:
Raw Data
Dataset
Processing
Decision

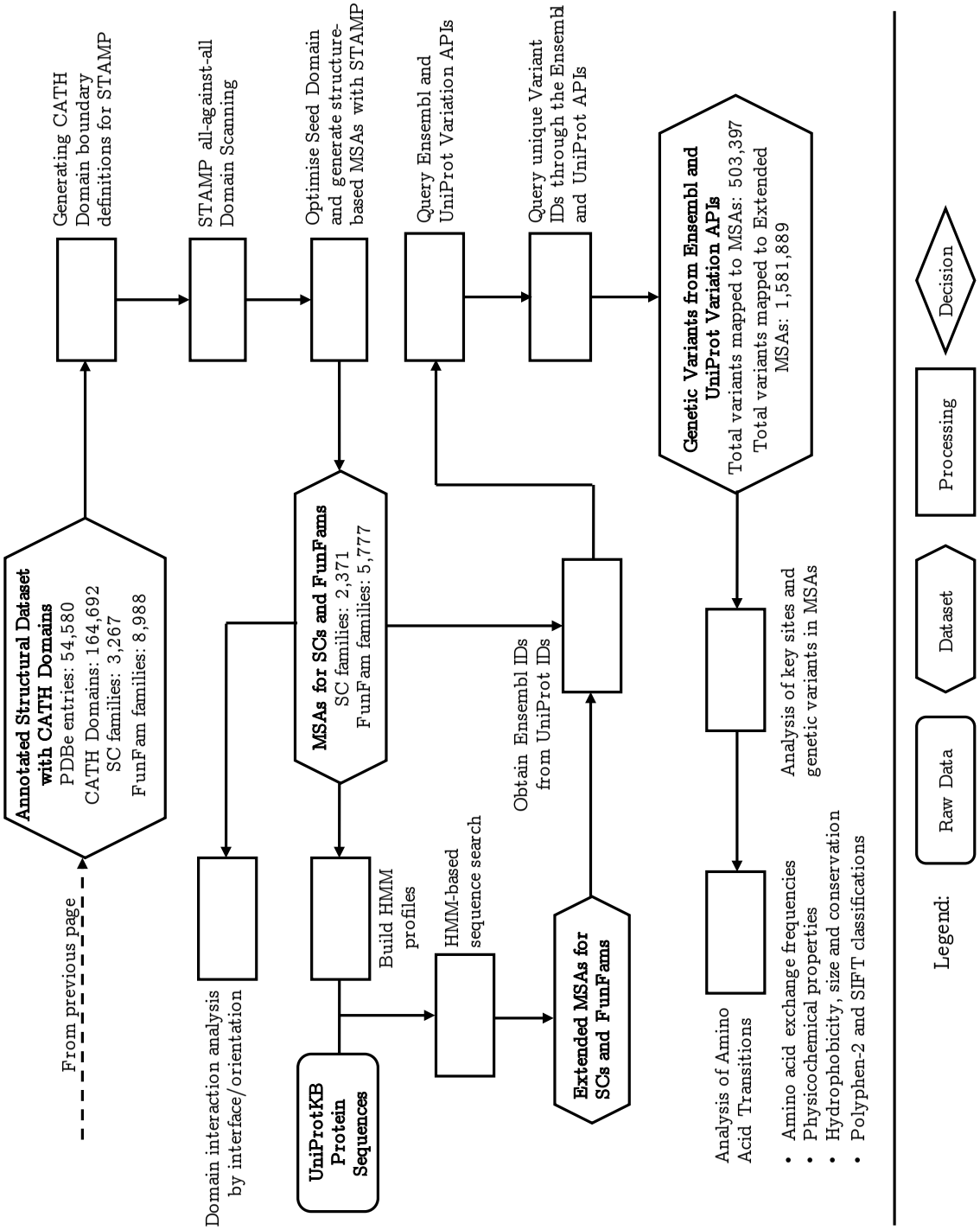

From previous page
Annotated Structural Dataset with CATH Domains
PDBe entries: 54,580
CATH Domains: 164,692
SC families: 3,267
FunFam families: 8,988
Generating CATH Domain boundary definitions for STAMP
STAMP all-against-all Domain Scanning
Optimise Seed Domain and generate structure-based MSAs with STAMP
MSAs for SCs and FunFams
SC families: 2,371
FunFam families: 5,777
Query Ensembl and UniProt Variation APIs
Query unique Variant IDs through the Ensembl and UniProt APIs
Genetic Variants from Ensembl and UniProt Variation APIs
Total variants mapped to MSAs: 503,397
Total variants mapped to Extended MSAs: 1,581,889
Domain interaction analysis by interface/orientation
Build HMM profiles
HMM-based sequence search
Obtain Ensembl IDs from UniProt IDs
UniProtKB Protein Sequences
Extended MSAs for SCs and FunFams
Analysis of key sites and genetic variants in MSAs
Analysis of Amino Acid Transitions
Amino acid exchange frequencies
Physicochemical properties
Hydrophobicity, size and conservation
Polyphen-2 and SIFT classifications
Legend:
Raw Data
Dataset
Processing
Decision

Figure 2.2: Flowchart overviewing the main processing steps performed by ProIntVar in order to allow the analysis of genetic variants in the context of CATH protein families, structure-based MSAs and interaction interfaces. Summary statistics are provided for the main datasets generated, as are the names of key tools/methods used by ProIntVar. For more information, please refer to the detailed description provided throughout Chapter 2.

A particular PDB protein structure usually contains multiple different entities commonly referred to as chains. Chains, as classically defined in the PDB-file format, used to describe the structure of proteins and their related ligands, as well as other molecule types, such as nucleic acids and sugars. The mmCIF file format introduced a new way of organising the various molecules found in the structures, in which chains are now viewed as entities. Entities can only be polymeric (protein, peptide or nucleic acid), non-polymeric (varied-size molecules and ions), or water (defining explicit water molecules). This means that a particular chain A in the PDB-format, which defined the atoms of a particular protein bound to a ligand, will now be described in the mmCIF as a polymeric entity A and a non-polymeric entity B. A naming cross-reference is also available for maintaining compatibility with legacy tools and databases.

In order to work with mmCIF structure files, several methods to read and analyse the structures were developed in this work to extend PDBx, a Python module (Westbrook, J., 2012), which is a lightweight read/write set of tools for PDBx/mmCIF files and its dictionaries. This module not only provides the ability to parse mmCIF files, but also to write out changes to the original files. This functionality is useful for writing biological assemblies from the original mmCIF files (as described in Section 2.3.4). Additional structure related annotations were downloaded from the PDBe API, particularly regarding the quality and validation of the structures

(Gore et al., 2012; Velankar et al., 2013, 2016).

Table 2.1 overviews the number of macromolecular structures determined by X-ray crystallography, NMR, cryoEM and other techniques in the PDBe. The majority of the structures (89%) were determined by X-ray crystallography. To define a uniform and high-quality structural dataset, protein structures were selected if they were solved by X-ray crystallography, the resolution was $\leq$3.0 Å, the overall quality was $\geq$10%, according to the PDBe API ('summary of global absolute percentiles' validation endpoint), and all atoms' coordinates were present (i.e. not only Cα atoms). The overall quality metric results from harmonic means of absolute percentiles of geometric metrics (e.g. Ramachandran angles, clash score, side chains clashes, etc.), reflections-based metrics (R-free, RSR (Real Space R-factor) Z-score) and both these metrics taken together (Gore et al., 2012; Velankar et al., 2016).

As shown in Table 2.2, the structural dataset contained 115,031 protein macromolecular structures (release of 13$^{\text{th}}$ January 2016) (Figure 2.2). A list of obsolete structures was also obtained from the PDBe FTP server so that datasets that reference those entries (e.g. CATH, see Section 2.3.8 for more information), could

Table 2.1: Breakdown of the experimental technique used to solve the structures available in the PDBe. The 3D structure of macromolecules was determined by X-ray crystallography, solution nuclear magnetic resonance (NMR), cryo-electron microscopy (cryoEM) and other methods. The structural dataset was analysed as of 13$^{\text{th}}$ of January 2016 and consists of a total of 115,031 structures.

Structure determination method	**Total count**
X-ray crystallography	102,622
Nuclear Magnetic Resonance (NMR)	11,169
Cryo-electron Microscopy (cryoEM)	934
Other techniques	306

Table 2.2: Overview of the macromolecular structures collected from the PDBe and analysed in ProIntVar. The total number of PDBe structures, as well as those for which there are SIFTS UniProtKB cross-reference mappings, is shown. Obsolete structures were also obtained from the PDBe for maximising compatibility with the domains defined in CATH. DSSP-based secondary structure annotation and solvent accessibility could not be calculated for all the structures due to PDB-format incompatibilities. A subset of protein structures was selected for further analysis in ProIntVar. The structural dataset was analysed as of 13th of January 2016.

Description	**Total count**
Structures in the PDBe	115,031
Obsolete structures in the PDBe, captured for compatibility with CATH	3,377
Structures for which SIFTS contained cross-references to the UniProtKB	109,316
Structures for which DSSP-based secondary structure and solvent accessibility could be calculated	104,774
Structures selected for subsequent analysis in ProIntVar	84,110

be traced to a new PDB entry, when available. Structures were kept if structure-sequence mapping was available as detailed in Section 2.3.3. This selection resulted in a working set of 84,110 structures (Table 2.1 and Figure 2.2). The selected structures map to 29,351 unique UniProtKB protein sequences, from which 5,483 are from human.

2.3.3 Mapping protein structure to protein sequence

Not all of the available macromolecular structures in the PDB are of proteins. As summarised in Figure 2.3, the sequence to structure mapping was performed with the aid of SIFTS (Structure Integration with Function, Taxonomy and Sequence) (Velankar et al., 2013). SIFTS is a semi-automated process for maintaining up-to-date cross-reference information between all protein chains in the PDBe and their UniProt entries (The UniProt Consortium, 2015). SIFTS was also used to obtain

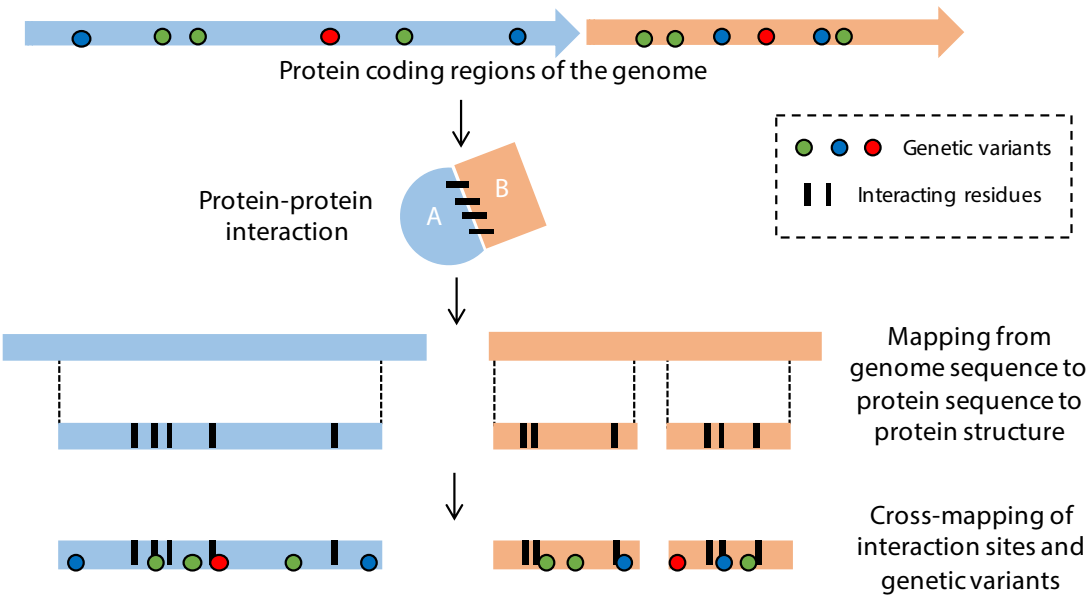


Figure 2.3: Multi-level genomic DNA sequence to protein sequence to protein structure mapping performed in ProIntVar. Protein structure to protein sequence mapping (PDB to UniProt) is performed by SIFTS. Protein sequence to genomic DNA sequence mapping (UniProt to Ensembl) is via UniProt ID mapping. As illustrated, this process allows for cross-mapping analysis of genetic variation in the context of protein structures and protein interactions.

cross-references between the PDBe and other biological databases, including Pfam (Finn et al., 2014a), SCOP (Lo Conte et al., 2000), CATH (Sillitoe et al., 2015), InterPro (Mitchell et al., 2015), and the NCBI taxonomy database (Federhen, 2012). PDB to UniProtKB mappings were available from SIFTS for 109,316 PDB entries.

2.3.4 Generating biological assemblies

As introduced in Section 1.4.2.5, the 3D coordinates that appear in the structures deposited in the PDBe are of the asymmetric unit (AsymUnit), the fraction of the unit cell that has no further crystallographic symmetry. Biological assemblies have been shown to provide a rich source of novel protein-protein interfaces that are not considered when using the AsymUnit coordinates deposited in the PDB (Jefferson et al., 2006; Bahadur et al., 2004; Bernauer et al., 2008). Thus, in the context of this

project, the structure of BioUnits are preferred, since these extended the number of interaction interfaces available for analysis.

Figure 2.4 overviews the process of generating biological assemblies from the asymmetric unit structures. Biological Units (BioUnits) are obtained by: A) removing crystal packing artefacts; and by B) applying symmetry operations (translation and rotation of atoms) to the various entities observed in the Asymmetric Unit, or by a combination of both. The process can both reduce the number of entities, as in the case of splitting the AsymUnit into multiple BioUnits (Figure 2.4 A), or increase the number of entities (Figure 2.4 B), for example by duplication (dimerisation in this example). For such cases that split the AsymUnit into two or more BioUnits containing different entities, all the BioUnits were analysed separately and in isolation.

Biological assembly annotations are defined in the mmCIF file and can have multiple evidence sources. Annotations are typically obtained from either the structure's depositors, from PDBe PISA/PQS (Krissinel and Henrick, 2007; Henrick and Thornton, 1998) software analysis, or a combination of both. By parsing and processing the mmCIF dictionaries that define the available biological assemblies and their related annotations, biological assemblies were generated by applying symmetry operations (rotation and translation of the residues' atoms) and/or removing crystal packing artefacts, achieved by splitting the AsymUnit into multiple BioUnits. The preferred biological assembly unit is defined and collected from the PDBe API (PDB endpoint). In this work, BioUnits were generated, as opposed to using precomputed structures in PDB-format available from the PDB. Doing so removes

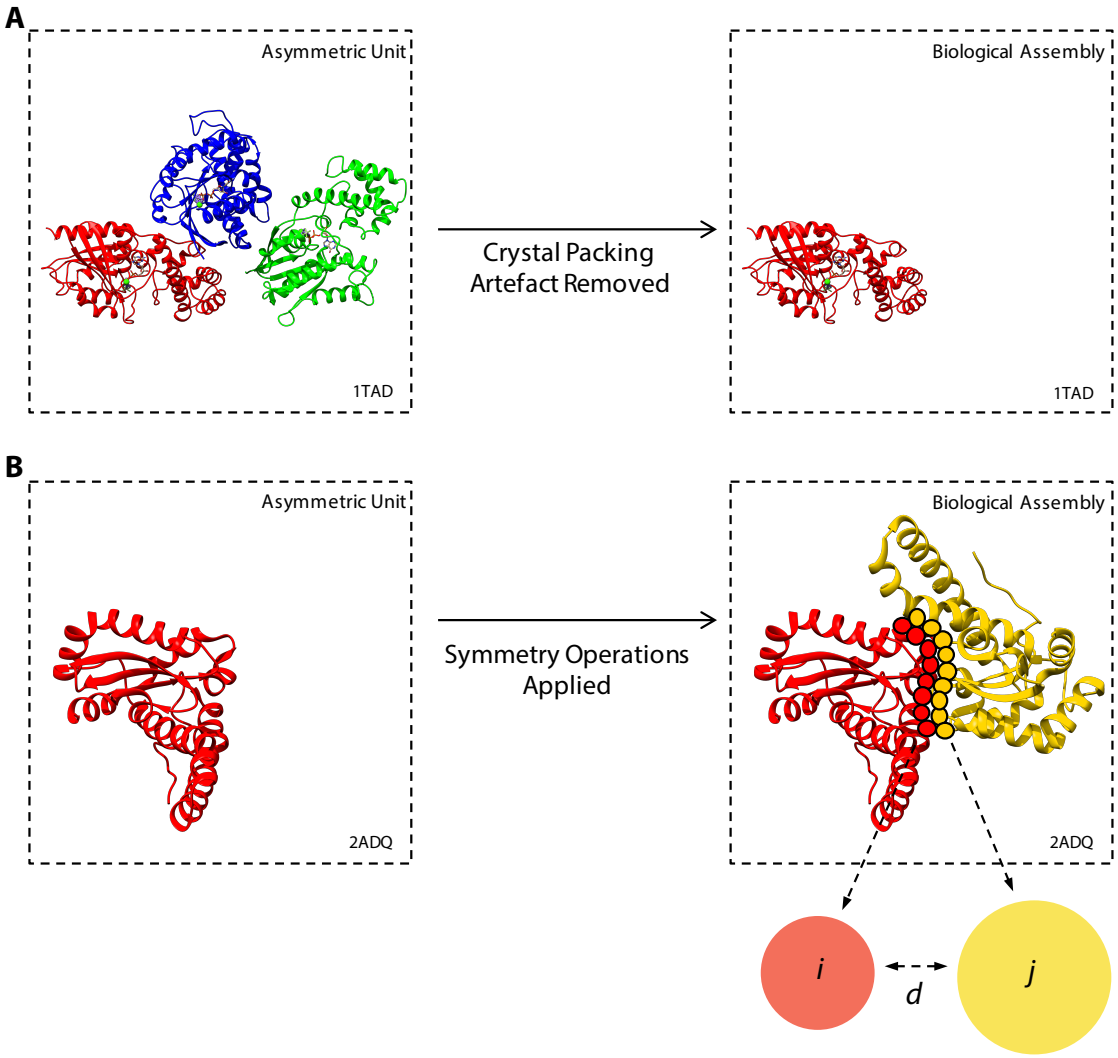


Figure 2.4: Diagram illustrating how biological assemblies and interaction interfaces are defined in ProIntVar. Biological Units are usually generated by: A) removing crystal packing artefacts; or by B) applying symmetry operations (translation and rotation of atoms) to the various entities observed in the Asymmetric Unit. Interaction interfaces are screened and defined as an atom-atom distance measure (d) which accounts for the atom's Var der Waals radii as well as the number of interacting interface residues.

one of the challenges in working with biological assemblies, namely the often non-straightforward mapping between the chain names in the AsymUnit structure, and their corresponding name in the BioUnit.

Table 2.3 overviews the generation of biological assemblies for the set of protein structures selected in ProIntVar. BioUnits were generated for 19,547 structures,

Table 2.3: Summary of the biological assemblies generated in ProIntVar. Biological assemblies (BioUnits) were calculated for all structures in the structural subset. The total number of entries for which the BioUnit is the same or differs from the asymmetric unit (AsymUnit) are shown, as are the BioUnits that result in higher-order oligomeric states (e.g. a dimeric AsymUnit produces a tetrameric BioUnit). The structural subset dataset consists of 84,110 structures and was analysed as of 13th of January 2016.

Description	**Total count**
Biological assembly is the same as the asymmetric unit	64,563
Biological assembly differs from the asymmetric unit	19,547
Biological assembly leads to a higher oligomeric state	19,199

while for the remaining 64,563 structures, the AsymUnit is believed to represent the natural oligomeric state of the protein (Table 2.3 and Figure 2.2). It is possible, though, that the resulting BioUnits are a subset (Figure 2.4 A) of the AsymUnit, for cases where the AsymUnit exhibits artificial or non-biological interactions, which might result from crystallisation artefacts (Krissinel and Henrick, 2007; Krissinel, 2011). These potentially arise from the experimental sample and crystallisation process and are described in more detail in Section 1.4.2.3. A total of 19,199 BioUnits calculated resulted in higher-order oligomeric structures that lead to an increase in the number of interactions available for structural analysis.

2.3.5 Defining protein interaction interfaces

Interactions between proteins and protein-ligands were determined based on atomic distance and performed as described in Jefferson et al., 2007b. Figure 2.4 B illustrates the process by which protein interactions and interacting residues are defined in ProIntVar. An all-against-all atom distance is calculated for all atom-atom pairs

that compose proteins and other entities observed in the structures. For speed purposes, the distance algorithm was improved to initially screen residues falling within 10 Å from each other. If this distance threshold is met, then the 3D Euclidean distance is captured and residue atoms were considered to interact if the distance $d = A_i + VDW_i - A_j + VDW_j$, where $A_i = (x_i, y_i, z_i)$ and $A_j = (x_j, y_j, z_j)$ are the relative positions of atom i and atom j in space, and where VDW is the Van der Waals radii (Tsai et al., 1999) of atoms i and j, respectively. A protein residue is classified as interacting when $d \leq 0.5$ Å and any of its atoms interact with any other atom of another protein residue or ligand. A minimum of 10 interacting residues (in total) was set as a cut-off to define protein-protein interactions. This threshold was chosen based on visual inspection of interaction interfaces and earlier work to remove false interactions arising from crystal packing artefacts (Jefferson et al., 2007b; Valdar and Thornton, 2001; Aloy et al., 2003). Since protein-ligand interactions can occur at the single residue-ligand, no threshold on the minimum number of interacting residues was set to define these interactions.

Protein interaction interfaces are held together by non-covalent forces including hydrogen bonds, ionic attractions (salt-bridges), Van der Waals forces, hydrophobic interactions, among others (Jones and Thornton, 1996; Janin and Chothia, 1990; Jones and Thornton, 1997; Hubbard and Argos, 1994; Argos, 1988). In addition to the atom distance measure, interaction types were computed based on standard residue (or atom) physiochemical properties. Intermolecular interaction types were calculated as defined in: Van der Waals (hydrophobic) (Voet and Voet, 2010); hydrogen bonds (McDonald and Thornton, 1994); salt-bridges (ionic interactions) (Kumar and Nussinov, 2002); aromatic-aromatic stacking (Burley and Petsko, 1985);

and disulphide bonds (Schmidt et al., 2006).

2.3.6 Removing crystallisation artefact ligands

As described in Section 1.4.2.4, it is common practice in structural biology to try various combinations of buffers and crystallisation compounds in order to solve the structures of proteins. This allows the optimal conditions to be identified, which produce high-quality diffracting crystals. Various molecules (e.g., Glycerol, Ethylene glycol) are often used as additives for solving the protein structures. Although this constitutes a crucial step in attempting to solve the structures, the resulting output might include molecules as well as interactions which are in many cases not biologically relevant or meaningful. Protein structure entities such as nucleic acids and a variety of ions, small molecules and sugars are herein referred as ligands. In this project, the BioLiP (Yang et al., 2012) dataset was employed to help filtering out such crystallisation artefacts (Figure 2.2). BioLiP (Yang et al., 2012) is a semi-manually curated database for high-quality, biologically relevant ligand-protein binding interactions that addresses this issue. The structure data are collected from the PDB, with biological insights mined from literature and other theme-specific databases. To construct a comprehensive and accurate database, BioLiP uses a composite automated and manual procedure for examining the biological relevance of every ligand observed in the structures. It enables cleaning of the protein structural dataset and allows a more meaningful analysis of protein-ligand interactions.

Table 2.4 shows the number of ligands observed in the structural dataset analysed in ProIntVar. From a total of 334,124 non-unique ligands, 244,757 were observed in the structures after crystal artefact filtering with BioLip (Yang et al., 2012). From

Table 2.4: Number of ligands observed in the structures after BioLiP filtering. A variety of ions, molecules, nucleic acids and peptides are observed in the structures. The structural subset dataset was analysed as of 13th of January 2016 and consists of 84,110 structures.

Description	**Total count**
Ions and molecules of varied-size/complexity	228,376
Nucleic acids	10,758
Peptides (protein chains $\leq$ 30 residues)	5,623

these, 228,376 corresponded to ions and other molecules (10,956 unique), 5,623 to small peptides (protein chains with length $\leq$30 residues) and 10,758 to nucleic acids.

The most common ligand types observed in the structures are nucleic acids, peptides, various ions such as Zn, Ca, Mg, Mn, Fe, Cu, sulfates, phosphates, and other molecules such as Heme-C, flavin adenine dinucleotide (FAD), adenosine diphosphate (ADP), adenosine triphosphate (ATP), flavin mononucleotide (FMN), α-d-mannose (MAN), among others. The scope of the protein-protein interaction (PPI) and protein-protein ligand (PLI) analysis in this project is currently limited to the complexes for which there are structures of directly bound protein partners in the PDB. Protein-protein and protein-ligand interaction interfaces were determined as described in Section 2.3.5. The total number of protein-protein and protein-ligand interactions observed in the filtered structural dataset was 161,222 and 242,809 in total, respectively.

2.3.7 Defining structural features and sites

Additional structural features and sites (or regions) (Goldman et al., 1998) were collected and analysed in ProIntVar (Figure 2.1 B). These include generating secondary structure annotations and computing the relative solvent accessibility of all

protein residues. Secondary structure was calculated for each protein structure using DSSP (Kabsch and Sander, 1983), which generates standardised secondary structure assignments from the 3D structures (Figure 2.2). DSSP identifies intra-backbone hydrogen bonds classifying eight types of secondary structure, including different types of helix, strands and turns or loops. For simplicity these were reduced to three main classification types: α-helix (H), β-sheet/strand (E) and turns or coil (T). The relative solvent accessibility (RSA) was calculated from DSSP's accessible surface values. For every residue in the structures, RSA was calculated as described by Ahmad et al., 2004:

$$RSA(x) = \frac{ASA}{SASA} \tag{2.1}$$

where the accessible surface area (ASA) is divided by the total solvent accessibility (SASA) of a standardised extended tripeptide Ala-X-Ala. X corresponds to the amino acid residue for which the RSA is being calculated (Ahmad et al., 2004). The total accessible surface values used for deriving RSA are shown in Table 2.5. Residues were classified according to the degree of exposure to the solvent as accessible (surface) where RSA $> 25\%$, partially exposed where $5\% <$ RSA $\leq 25\%$, or buried (core) where RSA $\leq 5\%$ (Cuff and Barton, 2000).

Table 2.5 shows a summary of amino acid properties, used in the various analysis performed in this work. Amino acid properties include: monoisotopic atomic mass (Knapp, 1996); average volume of buried residues, calculated from the surface area of the side chain (Zamyatnin, 1972); hydrophobicity (Fauchère et al., 1988); total accessible surface area (Ahmad et al., 2004); and frequency of occurrence (abundance) in the human proteome (de Beer et al., 2013).

Table 2.5: Summary of amino acid properties. Amino acid names, one-letter and three-letter codes as well as reference amino acid properties are provided.

Amino acid	**3-letter code**	**1-letter code**	**Surface area[a] (Å^2)**	**Volume[b] (Å^3)**	**Atomic mass[c] (Da)**	**Hydro-phobicity[d]**	**Human proteome[e] (%)**
Alanine	Ala	A	110.20	88.60	71.08	0.31	7.00
Arginine	Arg	R	229.00	173.40	103.14	1.54	5.63
Asparagine	Asn	N	146.40	114.10	128.13	-0.22	3.61
Aspartic acid	Asp	D	144.10	111.10	114.10	-0.60	4.77
Cysteine	Cys	C	140.40	108.50	57.05	0.00	2.29
Glutamic acid	Glu	E	174.70	138.40	147.18	1.79	7.13
Glutamine	Gln	Q	178.60	143.80	113.16	1.80	4.77
Glycine	Gly	G	78.70	60.10	137.14	0.13	6.54
Histidine	His	H	181.90	153.20	128.17	-0.99	2.64
Isoleucine	Ile	I	185.00	166.70	131.20	1.23	4.36
Leucine	Leu	L	183.10	166.70	113.16	1.70	10.00
Lysine	Lys	K	205.70	168.60	115.09	-0.77	5.80
Methionine	Met	M	200.10	162.90	129.12	-0.64	2.16
Phenylalanine	Phe	F	200.70	189.90	97.12	0.72	3.69
Proline	Pro	P	141.90	112.70	87.08	-0.04	6.23
Serine	Ser	S	117.20	89.00	156.19	-1.01	8.30
Threonine	Thr	T	138.70	116.10	101.11	0.26	5.29
Tryptophan	Trp	W	240.50	227.80	186.21	2.25	1.20
Tyrosine	Tyr	Y	213.70	193.60	99.13	1.22	2.66
Valine	Val	V	153.70	140.00	163.18	0.96	5.92

[a]Total accessible surface area (Ahmad et al., 2004); [b]average volume (Zamyatnin, 1972); [c]monoisotopic atomic mass (Knapp, 1996); [d]hydrophobicity (Fauchère et al., 1988); [e]frequency of occurrence in the human proteome (de Beer et al., 2013).

2.3.8 Collecting domain definitions from CATH

Protein structure domain definitions were obtained from CATH (Class, Architecture, Topology and Homologous Superfamily) (Sillitoe et al., 2015) and added to ProIntVar (Figure 2.1 C and Figure 2.2). CATH was used as the main protein domain classification here since the resource has a higher coverage in terms of annotated structures, when compared to SCOP. Additionally, CATH has been reliably maintained for around two decades and is still under active development.

The latest update to CATH v4.0.0 was made available on the 26th of March, 2013. As introduced in Section 1.5.1.1, CATH is a hierarchical classification of structural domains based on the analysis of their 3D structures. Since CATH domains are classified according to sequence, structural and functional similarity, domains clustered into the same Homologous Superfamily are expected to display an evolutionary relationship, with domains sharing significant structure, sequence and/or functional similarity and increasing levels of sequence similarity (45, 60, 95, 100% sequence identity). Structural clusters (SCs) and functional families (FunFams) are also defined below the Superfamily level in CATH, clustering together domains sharing structural and functional features, respectively (Lees et al., 2012; Sillitoe et al., 2013). As a result of the latest developments in the FunFam protocol in CATH (Das et al., 2015a), both structural domains as well as sequence predicted domains (based on FunFHMMer (Das et al., 2015b), refer to Section 1.5.1.1 for more information) are included in FunFams. In the context of this work, only the structure-based domains in FunFams were analysed since complementary methods were developed in this Ph.D. project for the analysis of protein structures and sequence predicted

domains (see Section 2.3.9 and Chapter 3 for more details).

In order to be able to work with as many CATH defined domains as possible, CATH domain definitions originally assigned to obsolete protein structures were transferred to the new PDB-remediated structures (based on PDBe identifier mapping) whenever possible. Additionally, to increase the confidence with which residues from a particular protein chain define a particular CATH domain, the subset of protein structures in ProIntVar was processed so that CATH domain boundaries were precisely mapped within the structures. This allowed annotation discrepancies to be removed from the domain sequence ranges obtained from CATH and the actual PDB residue-encoded sequences. This processing allowed a comprehensive set of CATH domain annotations to be defined that are seamlessly represented in the structural dataset in ProIntVar. A great deal of care was given to this processing step as it underpins the rest of the work reported in this Thesis.

Domain definitions were obtained from CATH and structural analysis of domain-domain interactions was performed (Figure 2.2). Table 2.6 summarises the contents of the CATH classification hierarchy. There are currently 235,858 domains defined, belonging to 4 classes, 40 architectures, 2,738 Homologous Superfamilies, 3,267 SCs and 8,988 FunFams, covering 69,058 proteins structures (<70% of the entire PDB). After filtering the structural dataset and mapping CATH domain boundaries to the structures, the number of Superfamilies (SPFs) in ProIntVar is 2,340, with 3,276 SCs and 8,988 FunFams. This corresponded to 164,692 unique CATH domains defined for the entire filtered structural dataset. The number of SCs and FunFams is reduced to 2,371 and 5,777, respectively, since families (SC or FunFam) that contain a single domain are removed, as at least two domains are needed for performing structural

alignment.

Table 2.6: Summary contents of the CATH structural classification hierarchy in the ProIntVar structural dataset. Domains are organised in CATH following the hierarchy from the top to the bottom, i.e. from Class to Architecture, to Topology, to Homologous Superfamily (SPF). SPF are further sub-grouped into structural clusters (SCs) and functional families (FunFams). The structural subset dataset was analysed as of 13th of January 2016 and consists of 84,110 structures.

Description	**Total count**
Class	4
Architecture	40
Topology	1,195
Homologous Superfamily (SPF)	2,340
Structural Clusters (SCs)	3,267
Functional Families (FunFams)	8,988
Unique Domains	164,692
Unique PDB Structures	54,580
Unique PDB Protein Chains	123,909

2.3.9 Generating multiple sequence alignments with STAMP for CATH SCs and FunFams

In order to investigate genetic variation in the context of protein domain families and to focus on key sites, such as interacting interfaces, structure-based multiple sequence alignments (MSAs) were generated for CATH SC and FunFam domain families. Structural alignment of domain structures and generation of a corresponding structure-based sequence alignment is essential for the analysis of protein structural and functional families in an evolutionary perspective. The comparison of protein structures can reveal distant evolutionary relationships that would not be detected by sequence information alone, or by the analysis of highly similar human

homologs. The analysis of variant sites throughout evolutionary history helps to infer those likely to affect key determinant sites for maintaining domain structural and functional integrity. Additionally, using structural alignments allows a deep domain structural coverage to be obtained by extending the MSAs with protein domain sequences of unknown 3D structure, and in this way expanding the structural coverage of the genetic variation/key sites analysis.

Since structure-based MSAs of all domains classified into SC and FunFam families were not readily available from CATH, multiple structure alignments were generated in this work (Figure 2.1 C and Figure 2.2). MSAs were generated by the STAMP (Structural Alignment of Multiple Proteins) program (Russell and Barton, 1992). A newly developed protocol for improving the structural alignment of domains was developed, as was a new protocol for extension of the STAMP alignments with similar sequences (Figure 2.2). The former involves fine tuning STAMP for the needs of aligning sets of structure domains, varying in number and structural diversity. The latter explores the power of Hidden Markov Model (HMM)-based sequence techniques to improve detection of similar sequences. Taking advantage of reliable STAMP multiple structure alignments, profile HMMs were built for each alignment, and sequence searches performed against the set of reference proteomes for which variation data is potentially available (see Section 2.3.11 for more details). Similar sequences were then re-aligned back to the STAMP structure alignments, in this way extending the structural alignments with additional protein sequences. The methods, as well as the corresponding results and discussion, are explored in detail in Chapter 3.

2.3.10 Analysis of CATH domain interactions

Protein domain interactions were analysed in the context of the CATH hierarchy for SCs and FunFams (Figure 2.1 C and Figure 2.2). Domain-domain interaction interfaces were defined as described in Section 2.3.5. In order to be able to analyse domain-domain interactions and to classify interactions by interaction interface, the relative orientation of the interacting domain pair was determined using the iRMSD method developed by Aloy et al., 2003, and further improved by Jefferson et al., 2007b. This method determines if two interacting domains are in the same orientation as another pair of interacting domains and thus if they are interacting with the same interaction interfaces. The iRMSD method uses the sequence alignments generated by STAMP (described in Section 2.3.9), to match equivalent positions from each separate partner of the interaction pairs to determine the transformation of one structure to another. Additional implementation details, as well as results and discussion on the analysis of domain-domain interactions, are further detailed in Chapter 5.

2.3.11 Collecting and organising sequence data

Protein sequence data was obtained from the UniProtKB database (The UniProt Consortium, 2015). Protein sequences, as well as annotations, were obtained either from the FTP server or UniProt's download/API web-pages. Sequences were obtained in FASTA format, whereas sequence annotations were downloaded in a variety of file formats including GFF (General Feature Format), UniProtKB file format and JSON (JavaScript Object Notation). Genomic sequence data, namely for

gene transcripts and transcript protein products were obtained from the Ensembl resource database (Chen et al., 2010; Flicek et al., 2013; Cunningham et al., 2015) through the Ensembl REST API, Variation API version 84 (Rios et al., 2010; Yates et al., 2014) (Figure 2.2). The version of the human genome build used throughout this work was the GRCh37 (Genome Reference Consortium Human Genome).

The sequences of complete reference proteomes (The UniProt Consortium, 2014) were downloaded from UniProtKB's FTP server. These sequence datasets were used in the STAMP extension protocol as described in Section 2.3.9 and Section 3.3.2, aiming at increasing the coverage of the STAMP structure-based alignments for genetic variants mapped to structurally/functionally similar protein sequences (Figure 2.2).

2.3.12 Mapping genetic variants to protein sequence

Genome sequence to protein sequence mapping is necessary to populate genomic features such as genetic variation entries onto a protein sequence (Figure 2.3). For this mapping, UniProt sequence IDs (accession identifiers) were converted to three main Ensembl stable identifiers, which include Gene, Transcript and Transcript Protein Product (referred to simply as Protein in Ensembl) (Chen et al., 2010; Flicek et al., 2013; Cunningham et al., 2015), using the Ensembl REST API (Yates et al., 2014) (Figure 2.2). Every variation entry was then examined for mismatches between the annotated variant sequence, the Ensembl transcript product sequence, as well as the actual UniProtKB sequence. When mapping errors were detected variants were dropped from the analysis.

To populate genomic features such as genetic variation entries onto a protein

sequence, protein sequence to DNA genomic coordinates was performed (Figure 2.3). This ID mapping is possible since the Ensembl database maintains a protocol for cross-referencing its stable IDs to other external sequence resources, such as the UniProt. Ensembl-UniProt cross-references are generated by the UniProt-Ensembl extensive cross-mapping collaboration effort, through projects such as Reference Proteomes (The UniProt Consortium, 2014), to map all protein sequences available in UniProtKB to the underlying genomic assemblies (The UniProt Consortium, 2014; Herrero et al., 2016).

2.3.13 Collecting and organising genetic variants

The analysis of genetic variation in this study focuses on variants that occur in the protein-coding regions of genes: non-synonymous (missense) single nucleotide polymorphisms (nsSNPs); stop gained/lost variants; frameshift variants; and insertion/deletion (non-frameshift) variants. Genetic variation data was obtained from the Ensembl database (Chen et al., 2010; Flicek et al., 2013; Cunningham et al., 2015) through the Ensembl REST API (Rios et al., 2010; Yates et al., 2014) as well as the UniProt Variation API endpoint (The UniProt Consortium, 2015) (Figure 2.2). For each entry, a series of annotations were stored including: source and original accession identifier; type of substitution; amino acid change, if any; clinical relevance, disease information and disease identifier (from OMIM (Online Mendelian Inheritance in Man [`http://omim.org/`] (Amberger et al., 2015)), if available); chromosome location; and Ensembl identifiers for Gene, Transcript and Protein; phenotypes and genotypes related information, if available; and Polyphen-2 (Adzhubei et al., 2010) and SIFT (Kumar et al., 2009) prediction scores. All of the raw data

from these data sources was pre-processed through the same pipeline, to increase consistency across variation sources. Table 2.7 summarises the original sources of genetic variation data obtained from the Ensembl and UniProt variation resources. Only human variation was obtained from the Ensembl and UniProt Variation APIs.

Careful pre-processing of the variation information was performed in ProIntVar such that overlapping information was maintained but key unique annotations were preserved and highlighted. Genetic variants were therefore classified according to the primary variation source: UniProt, Ensembl, dbSNP, 1kGP, ExAC, ESP, COSMIC, Humsavar, ClinVar, HapMap, PhenCode, DVGa, HGMD-public, and OMIM; and source curation type: manual or automated. Finally, variants were further organised based on the variant annotation as: germline nsSNPs, disease-associated variants (a subset of germline variants), and somatic mutations.

Genetic variation was collected and organised in ProIntVar for three main subsets of protein sequences: proteins in the PDB dataset not covered by CATH domain classification (PDB subset); CATH classified subset of the PDB (CATH subset), and lastly, hit sequences that extended the STAMP structural alignments (STAMP extended subset) (Section 3.3.2). The total variant counts shown were obtained for the full-length protein sequences, which are expected not to be covered by the entirety of the protein structures and protein domains. If a particular variation is mapped to multiple structures, only a single instance of the variant is added to the total count.

The number of uniquely mapped UniProtKB sequences for the PDB subset was 10,395, which resulted in 292,217 unique missense variants. 16,124 protein sequences in the CATH subset resulted in 503,397 unique missense variants. Lastly, 306,579

Table 2.7: Summary of genetic variation sources as collected from the Ensembl and UniProt variation APIs. Missense variants were collected from the UniProt API Variation endpoint (The UniProt Consortium, 2015) as well as the Ensembl REST API (Rios et al., 2010; Yates et al., 2014).

Source	From	Curation
dbSNP (Sherry et al., 2001)	Ensembl	Automated
1kGP (The 1000 Genomes Project Consortium, 2012)	Ensembl	Automated
COSMIC (Forbes et al., 2015)	Ensembl	Automated
HGMD-public (Stenson et al., 2014)	Ensembl	Automated
ClinVar (Landrum et al., 2016)	Ensembl	Automated
OMIM (Amberger et al., 2015)	Ensembl	Automated
PhenCode (Giardine et al., 2007)	Ensembl	Automated
HapMap (The International HapMap Consortium, 2007)	Ensembl	Automated
ExAC (Exome Aggregation Consortium (ExAC), Cambridge, MA, 2016)	UniProt	Automated
ESP (NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, 2016)	UniProt	Automated
Humsavar (Wu et al., 2006; Famiglietti et al., 2014)	UniProt	Manual

1kGP - 1000 Genomes Project; COSMIC - Catalogue of Somatic Mutations in Cancer; HGMD-public - Human Gene Mutation Database; ClinVar - Archive of Interpretations of Clinically Relevant Variants; PhenCode - Phenotypes for ENCODE; HapMap - Haplotype Map; ExAC - Exome Aggregation Consortium; ESP - Exome Sequencing Project.

sequences searched to extend the STAMP multiple structure alignments, resulted in 1,581,889 variants (Figure 2.2). This included coverage of an extra 8,069 protein sequences from human. Table 2.8 summarises variation sources and annotations as organised in ProIntVar. The total counts shown are not cumulative since the same unique variant can be simultaneously collected from multiple sources. Overall

analysis of genetic variants in the context of protein structures is further explored in Chapter 4 and Chapter 5.

2.3.14 Implementation and data analysis

The ProIntVar computational framework was developed in Python [`https://www.python.org/`] and related technologies. Data analysis and processing were performed with Python and R (R Core Development Team, 2016) [`https://www.r-project.org/`]. Plotting and visualisation were performed in R with the aid of ggplot2 [`http://ggplot2.org/`]. Figures of protein structures were generated using UCSF Chimera (Pettersen et al., 2004). MSAs and dendrograms were generated with the aid of Jalview (Waterhouse et al., 2009).

Statistical analysis was performed using the R core statistical package (R Core Development Team, 2016). Sample distributions were compared using the non-parametric Mann-Whitney-Wilcoxon test, or otherwise indicated in the text. A p-value lower than 0.05 was considered indicative of statistical significance. Sample correlation analysis (r) was calculated as the Pearson correlation coefficient. Amino acid variant counts and frequencies observed for the three variation classes were compared using Fisher's exact test implemented in the R package (R Core Development Team, 2016). Given the nature of Fisher's exact test, the p-values provided were not corrected for multiple comparisons.

Table 2.8: Overview of the genetic variants organised in ProIntVar. Genetic variants were collected from Ensembl and UniProt for the full-length UniProtKB proteins that are cross-mapped to three main structural subsets: PDB (not CATH-covered), CATH, and STAMP-extended. These are organised in ProIntVar according to the source, preprocessing (manual or automated), and annotation (germline nsSNPs, disease-associated variants, and somatic mutations). The total number of unique variants for which SIFT and Polyphen-2 prediction scores, as well as minor allele frequencies (MAF), could be obtained, are also shown.

Variation annotation	**PDB**	**CATH**	**STAMP extended**
dbSNP	46,581	173,966	389,326
1kGP	30,716	52,980	185,696
ExAC	171,736	296,590	974,684
ESP	38,201	66,347	222,594
HGMD-public	3,416	6,581	17,041
COSMIC	84,656	145,154	442,471
Humsavar	5,859	14,513	19,909
ClinVar	1,809	9,431	6,602
HapMap	428	2,011	3,430
PhenCode	1,215	4,544	1,811
OMIM	3,479	10,541	9,136
Manually curated	5,932	14,659	20,123
Automated	288,653	493,432	1,573,150
Germline nsSNPs	200,530	341,479	1,124,043
Somatic mutations	84,391	143,584	441,992
Disease variants	7,296	18,334	15,854
MAF	6,348	10,020	40,945
Polyphen-2	212,875	360,016	1,179,332
SIFT	142,022	235,433	784,352

dbSNP (Sherry et al., 2001), ExAC (Exome Aggregation Consortium (ExAC), Cambridge, MA, 2016), 1kGP (The 1000 Genomes Project Consortium, 2012), ESP (NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, 2016), HGMD-public (Stenson et al., 2014), COSMIC (Forbes et al., 2015), Humsavar (Wu et al., 2006; Famiglietti et al., 2014), ClinVar (Landrum et al., 2016), HapMap (The International HapMap Consortium, 2007), PhenCode (Giardine et al., 2007), and OMIM (Amberger et al., 2015).

2.3.15 ProIntVar web-server

A preliminary non-public web-server for ProIntVar has been developed to allow easy access to the results generated in this work. Similarly to the ProIntVar computational framework, the web-server was implemented in Python [`https://www.python.org/`]. The Flask [`http://flask.pocoo.org/`] Python web-framework was used to develop the ProIntVar web-server as it allows for easy development of RESTful APIs. The ProIntVar web-server gives access to multiple structural alignments, domain-domain iRMSD classifications as well as genetic variation information. Datasets can be browsed and downloaded from the web-server. A search toolbox allows common database IDs (including UniProtKB, CATH, Ensembl and Variation) to be searched for, providing a convenient and quick access point to lookup the analysis results. Results are visualised mainly through sortable tables and interactive plots. Protein structures with highlighted residues can also be visualised with LiteMol [`https://webchemdev.ncbr.muni.cz/LiteMol/`].

The ProIntVar web-server and database will be provided at `http://www.compbio.dundee.ac.uk/ProIntVar` at a later time. While the main purpose of ProIntVar is to allow access to the structural and sequence analysis performed in this work, it can be used as a generic tool for the analysis of variants in the protein structures. Access to the source code and documentation on how to deploy a local install of the software will be provided at a later stage, through the main project web-page.

2.4 Comparison to other tools/systems

Several tools/web-servers are available that support the analysis of genetic variation mapped to protein structure, with varying degrees of overlap and focus on structural analysis. Tools that allow analysis and visualisation of variants in structures include: SAAPdb (Cavallo and Martin, 2005), MutDB (Dantzer et al., 2005), ColiSNP (Kono et al., 2007), LS-SNP/PDB (Ryan et al., 2009), VnD (Yang et al., 2011), SNPdbe (Schaefer et al., 2012), and MSV3d (Luu et al., 2012). Cancer3D (Porta-Pardo et al., 2015) is a recently-developed web-server tool that focuses particularly on cancer somatic mutations from a variety of cancers and cell/tissue types. DMDM (Peterson et al., 2010) focuses on mapping variants on domains, while Structure-PPi (Vázquez et al., 2015) focuses on cancer mutations mapped onto protein-protein interaction interfaces.

The Single Amino Acid Polymorphism database (SAAPdb) (Hurst et al., 2009) links single nucleotide polymorphisms (SNPs) to phenotype alterations. SNP data is linked to a gene sequence, to determine whether the mutation occurred in a coding region; if so, the protein sequence and the mutated variations are displayed. Whenever possible, mutations are mapped onto protein structures, allowing the effect of the mutations on protein structure with the clinical phenotype to be investigated. LS-SNP/PDB (Ryan et al., 2009) is a resource for genome-wide annotation of human nsSNPs, which uses an automated, high-throughput pipeline for mapping the variants onto PDB structures and annotates several biologically relevant features. Although systems such as LS-SNP/PDB and SAAPdb initially seemed to be promising to be used as the main framework in this Thesis, they were not flexible enough

to allow thorough analysis of variation in the context of protein structure domain families and interaction interfaces. Additionally, none of these tools support the study of recently identified missense variants and newly solved structures.

2.5 Conclusions

ProIntVar is the main computational framework developed in this Ph.D. project. It encompasses all the tools and methods necessary for the analysis of genetic variants in the context of feature-rich protein structural data. A general overview of the datasets collected and organised in ProIntVar were overviewed in this Chapter. Additional method details and results are fully described in the following Chapters.

The main key features incorporated and highlighted in the ProIntVar computational framework are:

- ProIntVar provides a framework for the seamless analysis of sparse structural data, sequence data, interaction data and genetic variation data.
- It implements routines for generating biological assemblies, defining protein interactions, annotating additional structural features in sites and regions.
- It allows protein structure, protein sequence and genomic DNA sequence to be cross-mapped.
- It incorporates structure-based multiple sequence alignments for CATH structural clusters and functional families, which were further extended with similar protein sequences.

- It allows the comprehensive analysis of protein domain families and exploration of protein interfaces across different protein families from a structural perspective.

- It provides an integrative environment for the analysis of genetic variants in the context of proteins structures (atomic detail), as well as in the context of protein domain families (evolutionary detail).

- Finally, it annotates and organises genetic variation from a large set of sources and databases.

Chapter 3

Analysis of protein domain families

3.1 Summary

Structural alignment of similar protein structures and generation of a corresponding structure-based sequence alignment is essential for the analysis of protein structural and functional families. The comparison of protein structures can reveal distant evolutionary relationships that would not be detected by sequence information alone. This Chapter overviews the generation of MSAs from the alignment of protein structure domains organised in structural clusters (SCs) and functional families (FunFams) in CATH. MSAs are key to investigate genetic variation within and among species and are used in this project for the analysis of missense variants in protein interaction interfaces. To increase the structural coverage of the CATH domain classifications and enrich the analysis of genetic variants, new methods were developed

so that reliable MSAs are generated by the STAMP (Structural Alignment of Multiple Proteins) (Russell and Barton, 1992) alignment program. A protocol that uses profile Hidden Markov Model (HMM)-based methods (Eddy, 1998) was also developed so that similar protein sequences were annotated to extend the structure-based alignments.

3.2 Introduction

Protein structure comparison is a crucial step in studying the relationships between proteins, alluding to their functions and evolution. As protein structure often determines function, similarity in structure implies similarity in function (Todd et al., 1999). In fact, protein structure alignment has become an important technique for protein structure classification, protein function prediction, evolutionary relationship determination, molecular modelling and protein engineering (Murzin and Bateman, 1997; Abyzov and Ilyin, 2007; Launay and Simonson, 2008; Zhang et al., 2012; Koga et al., 2012), among others. Accurate alignment of the three-dimensional (3D) structure of proteins with near atomic-level resolution enables the detection and analysis of key conserved sites associated with protein function, such as catalytic sites (Valdar and Thornton, 2001; Halperin et al., 2004; Keskin et al., 2005). Besides, it provides important insights into functional mechanisms and the potential impact of genetic variation.

3.2.1 Structure Alignment Programs

As summarised in Table 3.1, various methods for multiple structure alignment have been developed over the last three decades (reviewed in Hasegawa and Holm, 2009). Among the most comprehensive methods are: STAMP (Russell and Barton, 1992), which is used throughout this work; DALI (Holm et al., 2008); TM-align (Zhang and Skolnick, 2005); SSM (Krissinel and Henrick, 2004); and Mustang (Konagurthu et al., 2006). Additionally, several methods originally developed for pairwise structural alignment: SSAP (Orengo and Taylor, 1996), CE (Shindyalov and Bourne, 2001); MAMMOTH (Lupyan et al., 2005); and FATCAT (Ye and Godzik, 2003); have been extended to allow multiple structure alignment. Although presenting important differences, all methods start by representing protein 3D structures in some coordinate-independent manner to make them comparable. This is typically achieved by constructing a series of matrices that encompass a variety of comparative metrics. One of the most commonly used metric is to take the pairwise 3D distances between some subset of the atoms (such as the Cα) in each structure (e.g. STAMP, CE, MAMMOTH and FATCAT). In fact, reducing the protein to a coarse metric such as structural fragments (SFs) (e.g. CE and SSAP) or secondary structure elements (SSEs) (e.g. SSM/PDBeFold), can also produce sensible residue-level alignments, despite the loss of information. This reduction is computationally important as the dimensionality of the matrices increases when the number of structures to be aligned increases. It is also common practice to apply dynamic programming (DP) techniques to the generated matrices (e.g. STAMP and TM-Align), and this allows a series of optimal local alignment paths to be determined,

Table 3.1: Summary features of several protein structure alignment programs.

Alignment Program	**Type**	**Class**	**Flexible**	**Year**
SSAP (Orengo and Taylor, 1996)	Multi	SSE	No	1989
STAMP (Russell and Barton, 1992)	Multi	Cα	No	1992
DALI (Holm et al., 2008)	Pair	SF	No	1993
CE (Shindyalov and Bourne, 2001)	Multi	Cα	No	2000
SSM (Krissinel and Henrick, 2004)	Multi	SSE	No	2003
FATCAT (Ye and Godzik, 2003)	Pair	Cα	Yes	2003
TM-align (Zhang and Skolnick, 2005)	Pair	Cα	No	2005
MAMMOTH (Lupyan et al., 2005)	Multi	Cα	No	2005
POSA (Li et al., 2014)	Multi	Cα	Yes	2005
Mustang (Konagurthu et al., 2006)	Multi	Cα	No	2006
Matt (Menke et al., 2008)	Multi	Cα	Yes	2008
FlexSnap (Salem et al., 2010)	Multi	SF	Yes	2010

Multi - Multiple structure alignment; Pair - Pairwise structure alignment; Cα - Backbone alpha-carbon; SSE - Secondary structure elements; SF - Structure fragments.

that are then summed to form a summary matrix. In some instances, a second round of DP is then performed on the summary matrix (e.g. SSAP). Although most methods are optimally tuned for rigid alignment of structures, methods such as FATCAT (Ye and Godzik, 2003) and Matt (Menke et al., 2008) have been developed to account for flexibility and structural rearrangements/inversions. Structural alignment techniques have allowed construction of all-to-all fold classification databases from the known protein structures in the PDB. Some examples of such databases are FSSP (Holm and Sander, 1996), CAMPASS (Sowdhamini et al., 1998), PDBeFold (Krissinel and Henrick, 2005), and CATH (Sillitoe et al., 2015).

3.2.2 Structural alignments by STAMP

STAMP (Russell and Barton, 1992) is a program that implements algorithms for comparison and superimposition of protein structures. STAMP makes extensive use of DP, least-squares fitting, and hierarchical cluster analysis techniques, as fully described in Russell and Barton, 1992.

Dynamic programming is a general technique which allows fast determination of the best path through a matrix containing a numerical comparison metric applied to all possible pairs of structure positions to be aligned. STAMP relies on the probability of residue structural equivalence measure (Pij) devised by Rossmann and Argos, 1976. The probability function consisting of distance and conformation terms (E_1 and E_2) is applied to the comparison of all pairs of residues and a DP path routine (modified Smith-Waterman algorithm) is used to identify the best set of equivalences between the pair. When aligning two domain structures A and B, the least-squares fitting takes a set of n Cα atoms (x, y, z) from A and n equivalent atoms from B and calculates the translation and rotation (transformation) that minimises the root mean square deviation (RMSD). This transformation can be applied to yield two new sets of coordinates for which calculation (and correction) of Pij values, the DP path-finding and the least squares fitting can be repeated iteratively until the two sets of residue coordinates, and the corresponding alignment, converge on a single solution. Hierarchical cluster analysis takes N domains and score measures calculated for the comparison of each of the $N(N-1)/2$ possible pairs of domains. Clustering returns a dendrogram that organises the domains according to their structural similarity.

STAMP provides two similarity measures to assess the quality of the structural alignments. Sc quantifies the global structural similarity between pairs or groups of domains, whereas Pij' provides a normalised measure of the confidence in the alignment of each domain's residue. The final structural similarity Sc score reported by STAMP results from a modified sum of Pij' scores corrected for the length of the pairwise alignments.

As illustrated in Figure 3.1, a good way to align a set of structure domains in STAMP is to start with a pre-computed set of domain transformations, which are selected from an initial all-against-all pairwise ($Pairwise$) domain scanning set (STAMP $Scan$). This allows STAMP to start aligning the domains based on a pre-superimposed set of 3D coordinates, instead of following a simple sequence-derived MSA. Multiple structure alignment then follows a procedure similar to tree-based multiple sequence alignment. Each possible pairwise comparison for the group of proteins to be aligned is performed. Structural similarity scores are used to derive a similarity matrix and corresponding dendrogram ($Treewise$). The dendrogram is then followed from the branches to the root superimposing structures in order of their similarity. MSAs are similarly generated by following the tree from the branches to the root.

A major advantage of STAMP is that it not only provides multiple structure-based MSAs and the corresponding superimpositions, but also provides a systematic and reproducible method for assessing the quality of such alignments. STAMP provides a series of alignments for domain structures thought to have a similar fold by following the systematically derived hierarchy of structural similarity. All steps in the alignment tree are preserved and can be individually assessed in order to improve

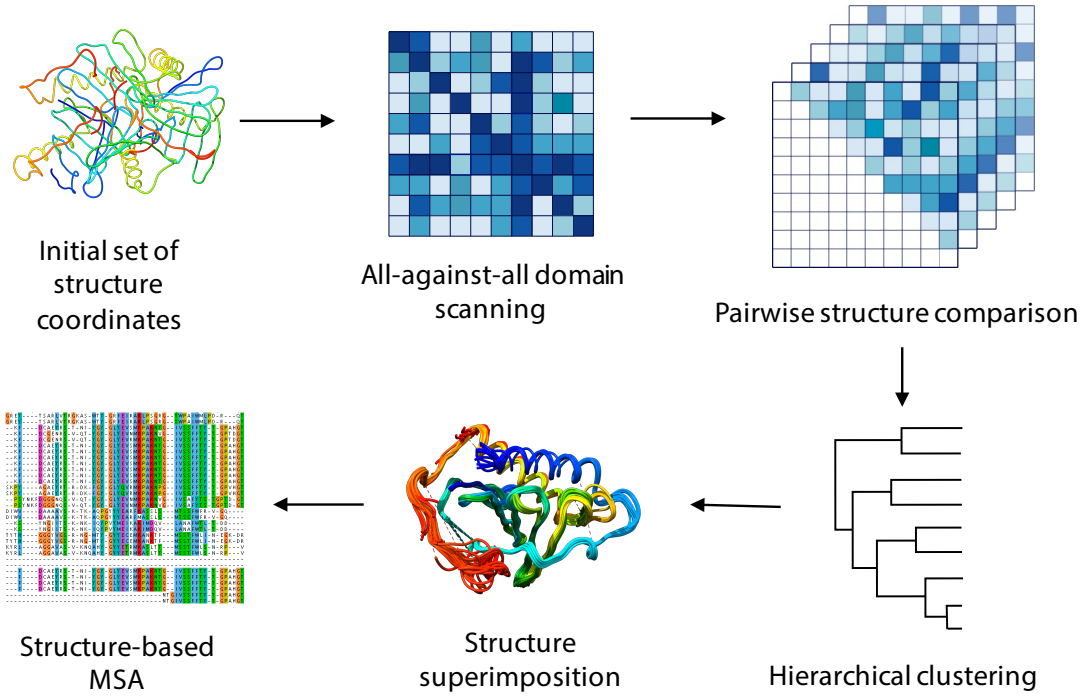


Figure 3.1: Schematic overview of the process performed by STAMP in order to generate multiple sequence alignments from the alignment of protein structures. The STAMP protocol works in five main stages: 1) all-against-all pairwise domain scanning; 2) selection of a seed domain representative set of coordinate transformations, which produce an overall higher scoring initial superimposition; 3) generation of a new superimposition and structure-derived tree based on the selected seed transformations; 4) further refinement of the superimposition found in 3) and creation of multiple sequence alignments and structural trees derived from the structural equivalences found; and 5) assignment of the reliability values to each region of the alignment. Darker blue colour in the heatmap indicates a higher Sc score. The example provided is for the Polo-box like domain (SPF:SC 3.30.1120.30:1).

the confidence of the final set of alignments. Additionally, STAMP assigns both the overall quality of alignment (Sc score) at each stage of the hierarchy and provides a confidence level for each aligned group of amino acid residues (Pij' score) within each alignment. This makes the quality metrics comparable among different protein families containing different numbers of domains and different sequence lengths.

Structure-based MSAs have been used extensively to improve the quality of sequence alignments (Shatsky et al., 2006), and have helped to benchmark sequence alignments derived without structure (Barton and Sternberg, 1987; Thompson et al.,

1999; Raghava et al., 2003). Deriving MSA profile statistics was an important development in computational biology that allowed the improvement of sequence similarity detection as implemented in popular iterative sequence search programs such as PSI-BLAST (Altschul et al., 1997). More recently, the SUPERFAMILY database (Gough et al., 2001) used HMMs to represent each protein family MSA in the SCOP database (Lo Conte et al., 2000). These profile HMMs were then used to identify sequence relatives to each SCOP family in a library of protein sequences. In a similar fashion, Gene3D has been developed to provide protein domain assignments based on the CATH hierarchy (Lees et al., 2012). Profile HMMs are similar to MSA profiles, but they contain position-specific probabilities for insertions and deletions, in addition to the amino acid frequencies per aligned column of an MSA. Profile HMMs have been shown to perform better than sequence profiles in detecting homologous proteins and generating hit MSAs (Eddy, 1998, 2011).

3.3 Methods

3.3.1 Generating structural alignments with STAMP

As introduced in Section 2.3.9, structure-based MSAs were generated by STAMP version 4.4.2 (Russell and Barton, 1992) for structure domains organised in SCs and FunFams in CATH (Sillitoe et al., 2015). In order to align CATH domains based on their structures, pre-processing of the generated biological assembly structures was performed as described in Section 2.3.8. Simple redundancy removal was performed by selecting only one copy of any domain which mapped to multiple chains in the biological assembly structures. Alignment of structures with STAMP (*Pairwise* and

Treewise) was performed using default parameters for E_1 and E_2, with an extended maximum domain length set to 1,000 amino acid residues. The two similarity measures provided by STAMP, Root Mean Square Deviation (RMSD) and Sc score, were used as proxies for the reliability of the structural alignments.

Figure 3.2 illustrates how an initial all-against-all domain STAMP *Scan* is performed (window size of 5 residues) to determine the best seed, i.e. the domain that produces a higher overall Sc scoring initial set of superimposed structures. This is performed since pairwise residue fitting with the Rossmann and Argos, 1976, function is improved if the protein structures being compared are approximately superimposed initially. For each pair of domains A and B, both A to B and B to A STAMP scans are performed and the initial superimposition problem is solved by attempting more than one initial fit between A and B. The best scoring domain is selected if it can generate better initial superimpositions with as many other domains in the family as possible. From the initial set of transformations obtained by superimposition with the seed domain, a survey over the domain pairs that are not optimally aligned is performed in order to test whether the reverse transformation (i.e. B to A instead of A to B) should be selected instead (Figure 3.2). The STAMP $Pickframe$ program was used to reverse the set of transformations generated for the seed domain by the initial STAMP scan, for those domain-domain pair scans which the reverse transformation is preferred.

Percentage sequence identity (PID) reported throughout this Chapter is provided by STAMP, and is defined from structure comparison as the percentage residue identity within structurally conserved (equivalent) regions (SCRs) (stretches of $\geq$3 positions with $Pij' \geq 6.0$). Root mean square deviation (RMSD) scores are iteratively

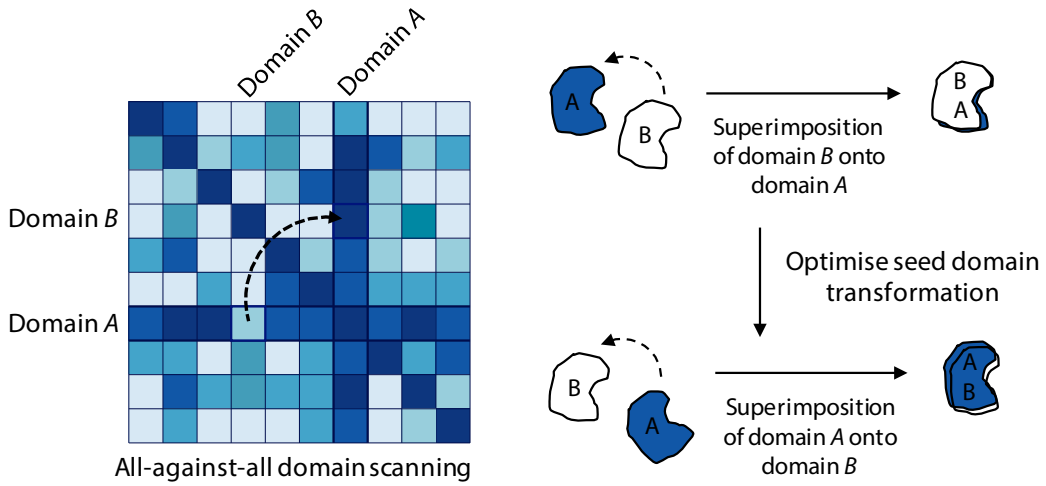


Figure 3.2: Improving the reliability of structure-based STAMP MSAs by optimising the set of transformations obtained for the seed domain. After performing an all-against-all domain STAMP scan, the higher *Sc*-scored seed domain is selected. In order to further improve the quality of the superimpositions and MSAs generated by STAMP alignment, the initial set of transformations is optimised by reversing the coordinate transformations obtained for low scoring superimpositions. Darker blue colour in the heatmap indicates higher *Sc* score.

calculated by the least-squares fitting only for the subset of structurally equivalent Cα atom positions. RMSD scores are similarly provided by STAMP.

3.3.2 Extending STAMP structural alignments with HMMs

As summarised in Figure 3.3, profile HMMs were generated for each multiple structure alignment obtained for SCs and FunFams. MSAs were converted to Stockholm file format and profile HMMs generated with the HMMER3 *hmmbuild* tool (Eddy, 2011). These profiles were then used to search a set of protein sequences from complete proteomes (see Section 2.3.11). A profile HMM sequence search was performed with the HMMER3 *hmmsearch* tool (Eddy, 2011) using inclusion/reporting significance thresholds *E-value* = 1 x 10^{-4}. MSAs containing the hit sequences were obtained with the same tool and merged with the full structure-based MSA (Figure

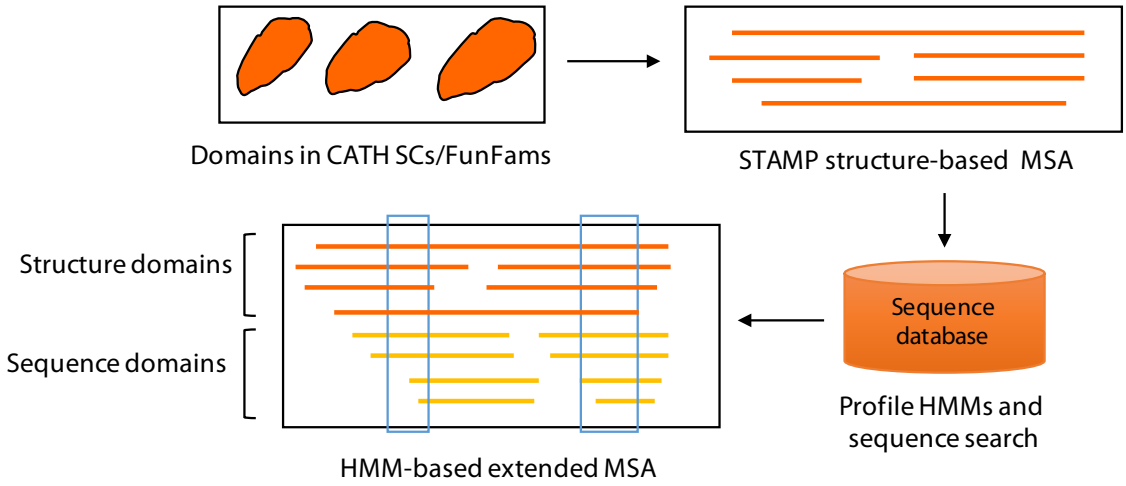


Figure 3.3: General overview of the protocol developed to extend STAMP structural alignments with similar protein sequences. CATH SCs/FunFams are structurally aligned with STAMP and MSAs generated. HMMER 3 tools (Eddy, 2011) are used to generate profile HMMs from these MSAs and used to search a sequence database containing protein sequences from complete reference proteomes. Finally, the original structural alignments are extended with the search hit sequences (in yellow) that pass the *E-value* = 1 x 10^{-4} threshold. Structurally equivalent positions (SCRs) are highlighted (blue square boxes) in the final extended MSA.

3.3). Whenever the hit sequences did not cover the entirety of the structure-based alignment, these were extended with gaps in order to retain the length of the original structure-based alignment.

3.4 Results and discussion

3.4.1 Improving the quality of the structural alignments generated by STAMP

Structure-based multiple protein sequence alignments (MSAs) were generated for both CATH SCs and FunFams. Although seed MSAs were available for both SC

and FunFams from the CATH database, working with these proved extremely complicated. This was mainly due to inconsistencies between the CATH domain definitions and the domain sequences depicted in the final seed MSAs. Structure-based MSAs were therefore generated by STAMP (Russell and Barton, 1992), due to our extensive knowledge of its features and the recognised quality of the resulting alignments (Raghava et al., 2003). STAMP follows a similarity hierarchy, and the obtained transformations and corresponding MSAs are output at each node of the dendrogram so that each sub-alignment may be examined separately. This is a clear advantage which allows for a fine structural analysis at increasing levels of structural similarity or diversity. STAMP assumes overall topological similarity within the protein families to be aligned, and would not be able in principle to superimpose/align structures with common secondary structures in similar orientations, but different connectivity or topologies. This requirement was fulfilled since both SCs and FunFams are expected to be structurally similar as they are clustered under the same Topology and Homologous Superfamily in CATH. Alternative structure alignment programs (reviewed in Table 3.1), were also initially considered but unfortunately were not suitable for generating the MSAs needed in this work. In addition to programs that only perform pairwise structural alignment (e.g. DALI (Holm et al., 2008) and TM-align (Zhang and Skolnick, 2005)), some were no longer available for use (e.g. SSAP (Orengo and Taylor, 1996)); others were only available through webservers which were not practical for aligning such a high number of domains (e.g. FlexSnap (Salem et al., 2010) and POSA (Li et al., 2014)); or imposed limitations on the number of domains to be aligned (e.g. MAMMOTH (Lupyan et al., 2005) and Mustang (Konagurthu et al., 2006)).

As described in Section 3.3.1, STAMP domain scanning was used to generate an initial set of superimposed coordinates. This is achieved by scanning all-against-all domains to be structurally aligned which belong to a particular SC or FunFam. A seed domain is selected based on the overall Sc score. Transformations obtained for the seed domain can be further optimised. Therefore these have been reversed in order to generate the best-starting transformation set possible (Figure 3.2). Because the difference in Sc scores can be small, only when at least a 0.25 Sc score difference is observed, the reverse transformation is preferred. This approach was applied to the entirety of CATH SCs and FunFams and led to a relative improvement of the initial set of transformations, for 46% of those domain families.

As shown in Section 2.3.8, the number CATH FunFams is more than double the number of CATH SCs (Table 2.6). This results in the number of domains clustered into SCs to be higher than those classified into FunFams. Figure 3.4 overviews five metrics used to compare the contents of STAMP MSAs generated for CATH SCs and FunFams. The MSAs have on average 142 residues in length (AL) ($21 \leq$ AL ≤ 898). Although the distribution of MSA lengths is comparable, MSAs generated for CATH SCs are slightly longer than those generated for CATH FunFams (Figure 3.4 A), which results from the higher diversity and number of domains aligned in SCs, as well as the resulting insertion of gaps. The distribution of the length of the shortest domain sequence (LSS) is similar for the 72% CATH domains, which are simultaneously clustered into SCs and FunFams (Figure 3.4 B). The distribution of the smallest percentage sequence identity (PID) scores observed for members of the various CATH SCs and FunFams, as calculated by STAMP, is also provided in Figure 3.4 C. PID scores range from 0 to 100 and a broader distribution of PID

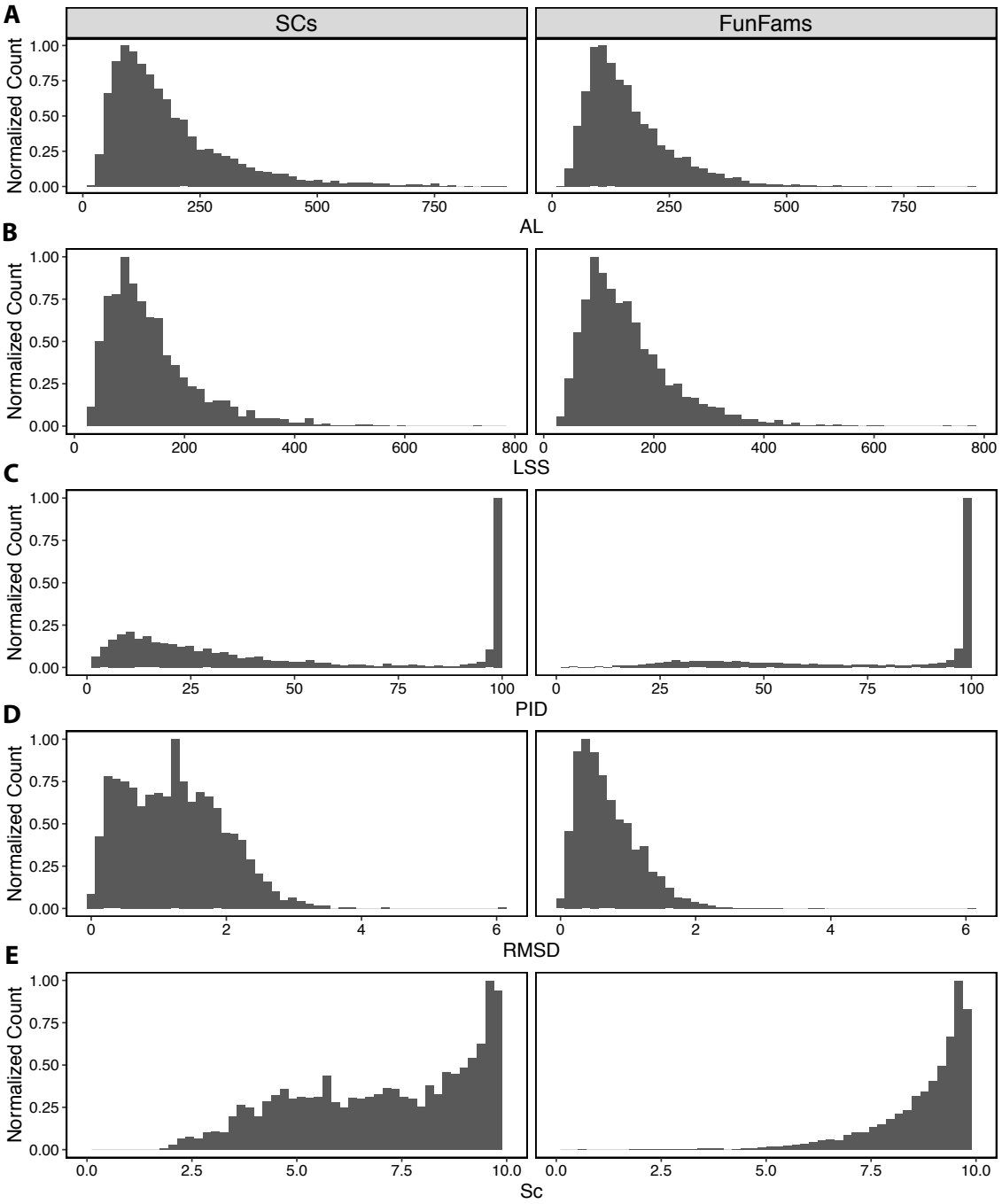


Figure 3.4: Overview of various STAMP alignment metrics for the comparison between SCs and FunFams. Histograms are provided for comparing: A) alignment length (AL); B) the length of the shortest domain sequence (LSS); C) the smallest percentage sequence identity (PID) observed; D) the Root Mean Square Deviation (RMSD); and E) STAMP *Sc* scores. Histogram binning was performed using normalised counts. The number of SCs and FunFams is 2,371 and 5,777, respectively.

scores is observed for the SCs, when compared to the FunFams, which distribute closer to a higher PID. This suggests that clustering of domains on the basis of function in CATH FunFams results in lower structural diversity.

Figure 3.4 D shows the distribution of RMSD scores obtained from STAMP by the superimposition and alignment of the domains in CATH SCs and FunFams. A Pij' of 6.0 was used to define structurally conserved (equivalent) positions in the structure superimpositions (and alignments). This threshold was found to yield good superimpositions and alignments (Russell and Barton, 1992), where lower Pij' scores generally result in poor fit. Higher Pij' scores generally result in too few structure equivalences. According to analysis of various protein families performed with STAMP (Russell and Barton, 1992), stretches of three or more aligned positions with Pij' values greater than 6.0 correspond to true topological equivalences, values between 4.0 and 6.0 are equivalent >50% of the time, and values less than 4.0 are more often not equivalent. Stretches of residues having $Pij' \geq 6.0$ generally correspond to regions of conserved secondary structure within a family of structures being compared. RMSD scores are obtained for structurally equivalent positions ($Pij' \geq 6.0$) and range from 0.0 to 2.0. A broader distribution of RMSD scores, ranging from 0.0 to 3.5, is observed for SCs. Lastly, Figure 3.4 E shows a wider dispersion of STAMP Sc quality scores for SCs when compared to those observed for FunFams, which are generally highly scored. Again according to the original STAMP analysis (Russell and Barton, 1992), alignments having a structural similarity score Sc between 5.5 and 10.0 display a high degree of structural similarity, which suggests a functional and/or evolutionary relationship. Scores between 2.5 and 5.5 correspond to more distantly related structures and do not always imply a

functional or evolutionary relationship. Scores <2.0 generally indicate little overall structural similarity.

3.4.2 Analysis of structurally conserved regions in the structural alignments

A complementary metric to analyse the reliability of the STAMP structure-based MSAs is to calculate the number of structurally equivalent positions (NEP) divided by the LSS. This corresponds to the number of residues that are considered to be structurally conserved in STAMP ($Pij' \geq 6.0$) relative to the smallest number of residue positions that could be equivalently aligned. Figure 3.5 shows the distribution of STAMP structurally conserved residues (NEP / LSS) when comparing CATH SCs and FunFams. An overall higher number of STAMP structurally conserved positions is observed for MSAs generated for CATH FunFams than for SCs.

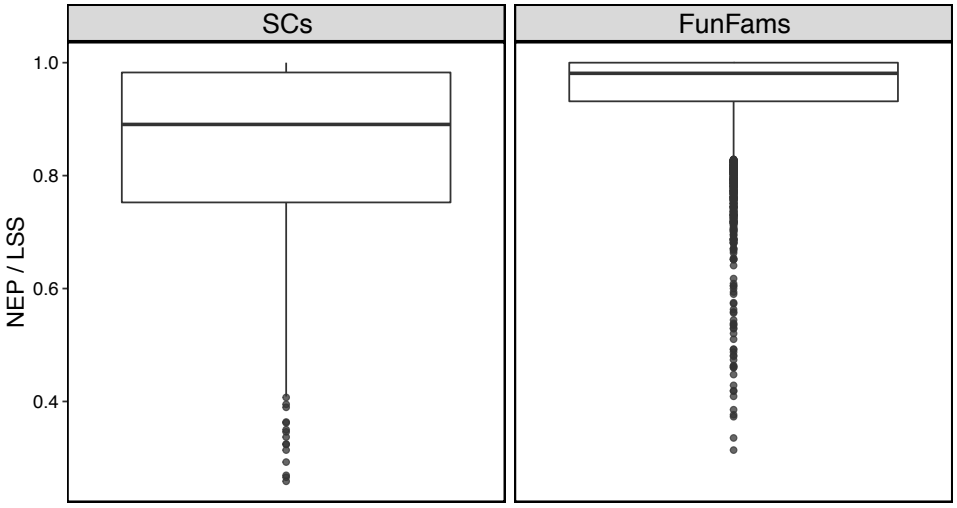


Figure 3.5: Box-plot showing the distribution of structurally conserved positions in the CATH SC- and FunFam-based STAMP MSAs. The number of STAMP structurally equivalent positions (NEP) is divided by the length of the shortest domain sequence (LSS). The distribution of NEP/LSS is provided for both CATH SCs and FunFams.

This results from the lower number of domains per protein family in FunFams, as well as the fact that FunFams display overall lower PID and higher STAMP *Sc* scores (Figure 3.4).

Figure 3.6 shows the correlation between the distribution of structurally conserved positions and the STAMP RMSD, *Sc* scores and PID, obtained for the STAMP structural alignments generate for SCs and FunFams. A high Pearson correlation coefficient is observed for the comparison of both NEP / LSS and RMSD (Figure 3.6 A) (r = -0.75 and r = -0.63, for SC and FunFam, respectively), as well as NEP / LSS and *Sc* (Figure 3.6 B) (r = 0.90 and r = 0.86, for SC and FunFam, respectively). Following the trend observed for the RMSD and *Sc* scores (Figure 3.4 D and E), the score distribution is tighter for FunFam-based alignments when compared to those obtained for SCs. A weaker correlation is observed for the comparison of NEP / LSS and PID (Figure 3.6 C) (r = 0.73 and r = 0.59, for SC and FunFam, respectively).

Figure 3.7 shows the trend of NEP / LSS, *Sc* and PID, for increasing numbers of domains per SC/FunFam family. The number of domains that compose each SC/FunFam in the analysis shows that the larger the number of domains, the lower the number of structurally equivalent positions in the resulting superimpositions/MSAs (Figure 3.7 A). The same profile is observed for both *Sc* and PID, where a higher number of domains per SC/FunFam results in lower detection of structural equivalence (*Sc*) (Figure 3.7 B) and lower PID (Figure 3.7 C). The effect is observed more markedly for SC than to FunFams, since, as expected, the number of domains and the structural diversity is higher for SCs than for FunFams.

Figure 3.8 shows three examples of SC/FunFam protein families superimposed

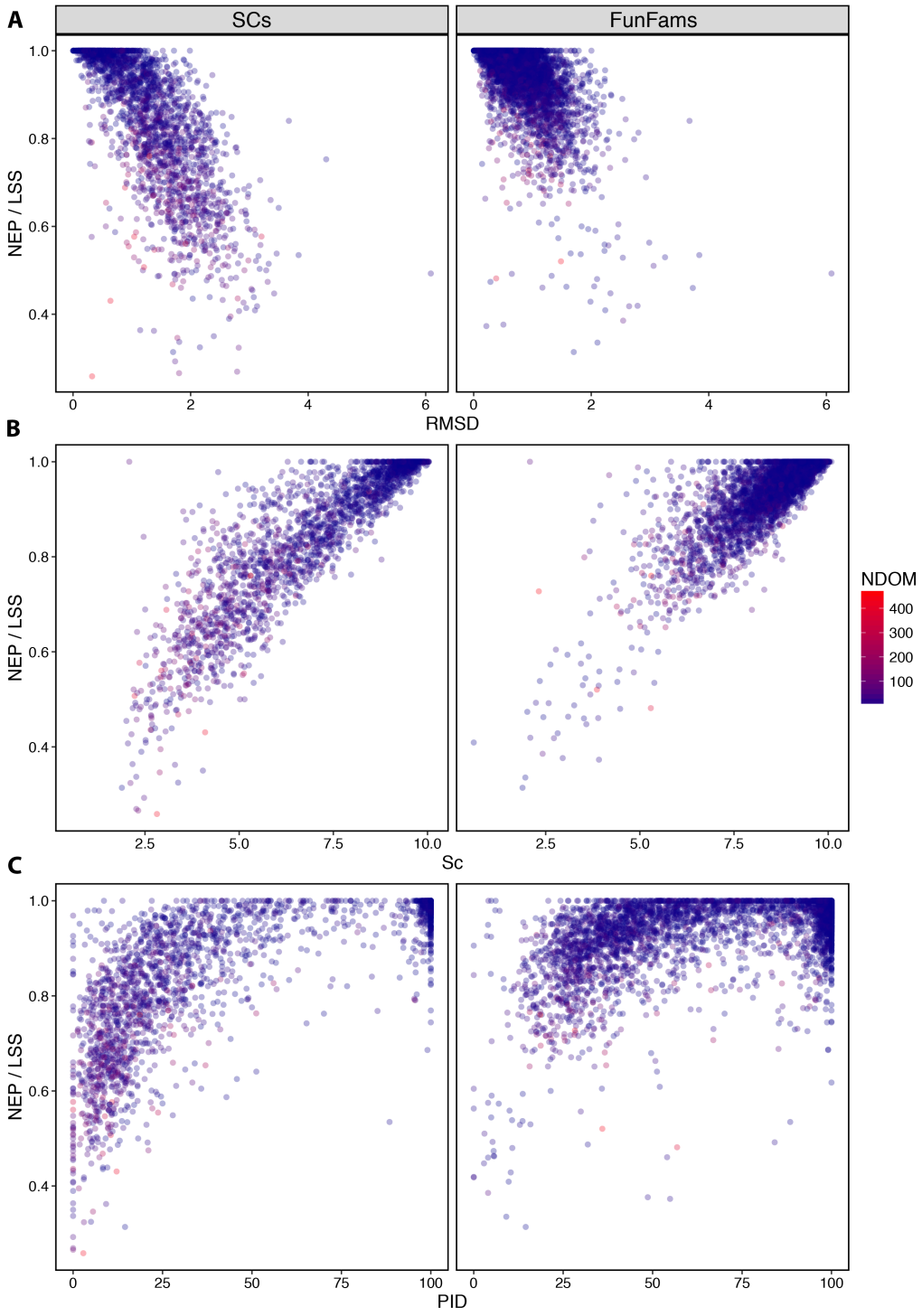


Figure 3.6: Assessment of STAMP alignment reliability with RMSD, *Sc* and PID, for SCs and FunFams. Pearson correlation coefficients A) of -0.76 and -0.63 were obtained for the number of equivalent positions (NEP) divided by the length of the shortest domain sequence (LSS) *versus* Root Mean Square Deviation (RMSD); B) 0.90 and 0.86 for NEP/LSS *versus* STAMP *Sc* score; and C) 0.73 and 0.59 for NEP/LSS *versus* the smallest percentage identity (PID) for SCs and FunFams, respectively. NDOM corresponds to the number of available domains per family.

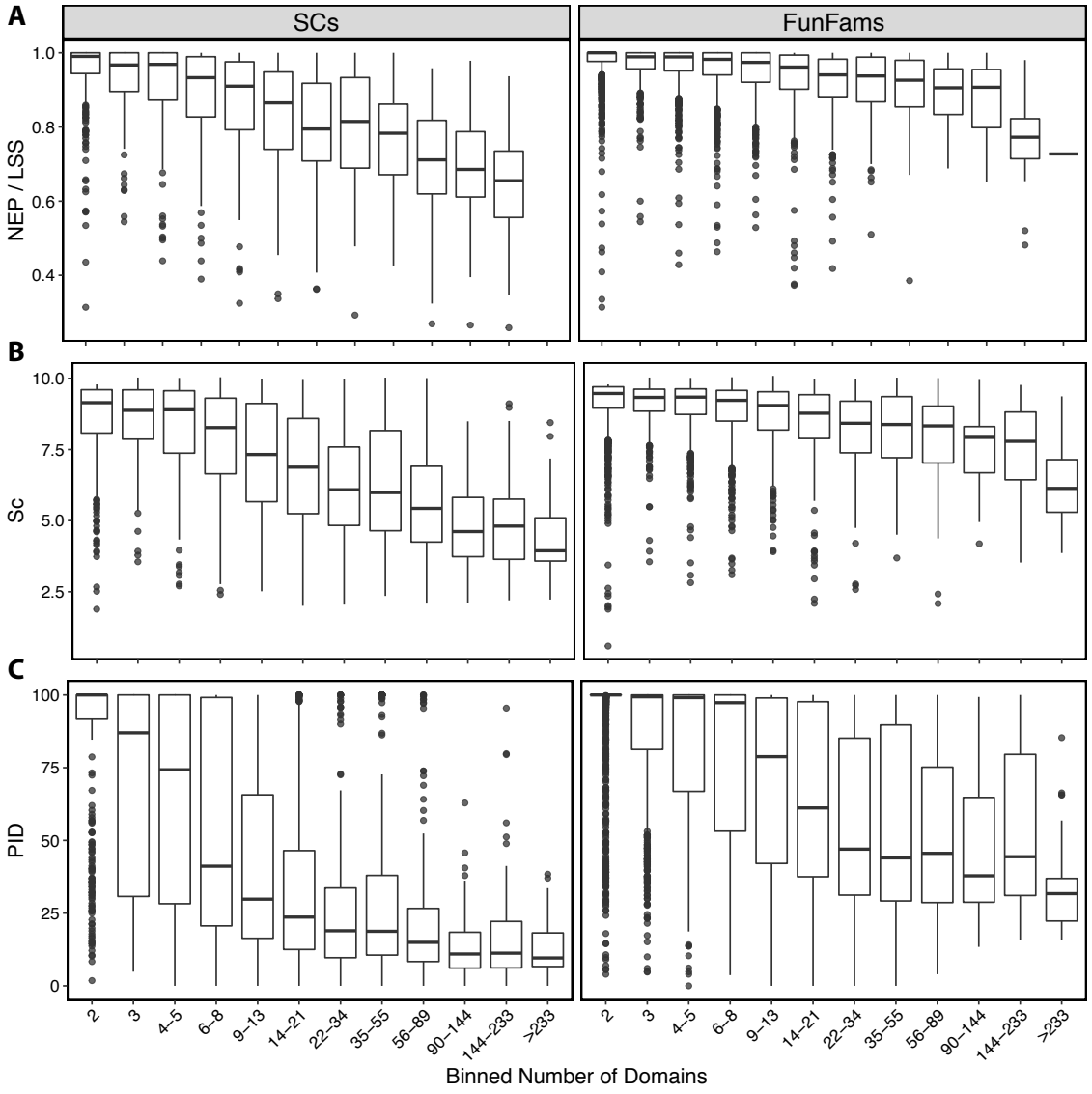


Figure 3.7: Assessment of STAMP alignment reliability comparing NEP/LSS, *Sc* and PID, over the number of domains, for SCs and FunFams. A) Number of equivalent positions (NEP) divided by the length of the shortest domain sequence (LSS); B) STAMP *Sc* score; and C) the smallest percentage identity (PID) for SCs and FunFams, respectively.

by STAMP, which despite having a low STAMP *Sc* score, show that the backbone residues are overall well superimposed. Figure 3.8 A, shows the structure superimposition of Actin-binding protein, T-fimbrin (CATH SPF:SC 1.10.418.10:1). This example highlights the case where a high RMSD of 3.31 is obtained despite good overall backbone superimposition (*Sc* of 3.07), as the result of relative low number

of sites considered to be structurally equivalent (conserved) (NEP/ LSS = 0.47). Regardless of good overall superimposition, the lowest PID observed within this SC is 0%.

Figure 3.8 B shows the structure superimposition generated for the Putative genome polyprotein (SPF:SC 2.60.120.20:5), which contains 45 domain members and a low overall PID of 17%. Multiple structure alignment of this SC by STAMP results in a Sc score of only 2.49. Despite this score, the RMSD score obtained for the structurally equivalent positions is 0.53, indicating that many residues in the backbone of the structure domains are well aligned. Nevertheless, the distribution of NEP divided by the LSS shows that in fact only 35% of the sites to be aligned are considered structurally equivalent ($Pij' \geq 6.0$, and hence used for RMSD calculation) by STAMP.

Finally, Figure 3.8 C shows the superimposition of domains in the S-layer homology domain ribonuclease family (CATH:FunFam 3.10.450.30:600). Despite showing related protein function, domains belonging to this FunFam display structural diversity (and sequence diversity, lowest PID = 4%). The STAMP Sc obtained for this FunFam was 2.41, which resulted in an RMSD of 2.54, for 38% of sites which were defined as structurally conserved by STAMP.

3.4.3 Increasing the structural coverage of the STAMP alignments for CATH SCs and FunFams

STAMP structural alignments were expanded with similar protein sequences as described in Section 3.3.2. Figure 3.9 shows the increase in the number of sequences

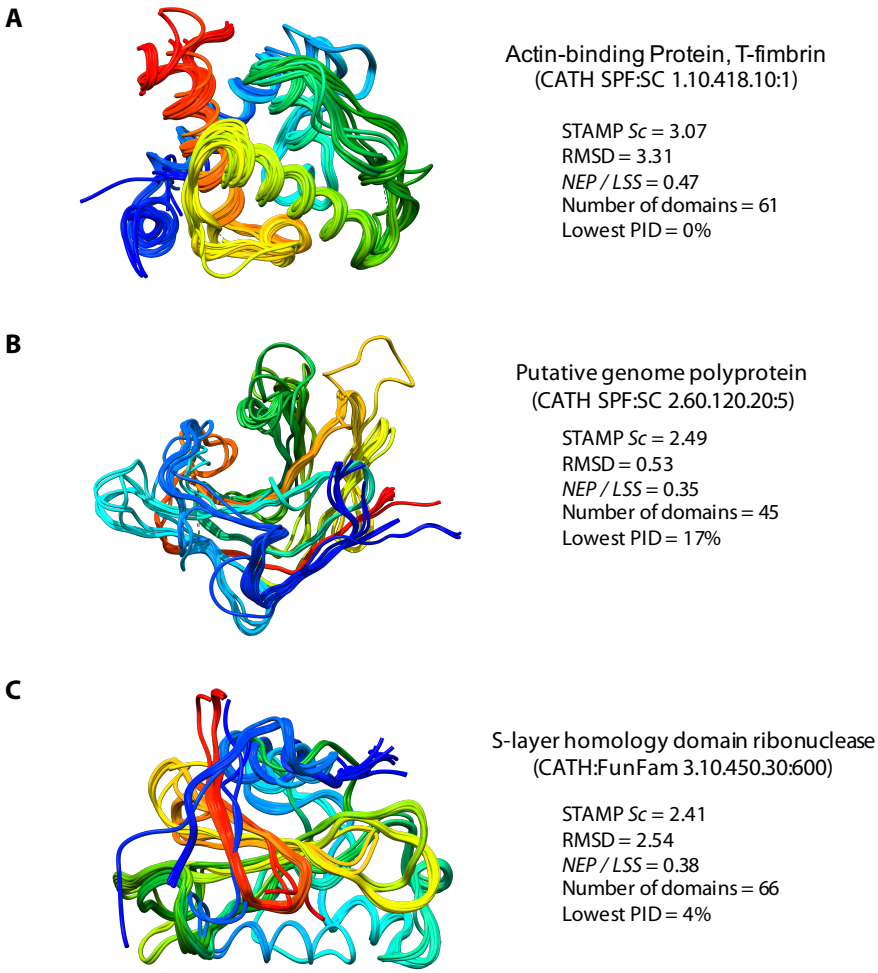


Figure 3.8: Example of low STAMP *Sc* scoring structure superimpositions obtained for CATH SCs and FunFams. The superimposition and SC/FunFam summary information is shown for: A) Actin-binding protein, T-fimbrin (CATH SPF:SC 1.10.418.10:1), which contains 61 domains; B) Putative genome polyprotein (CATH SPF:SC 2.60.120.20:5), which contains 45 protein domains; and finally C) for S-layer homology domain ribonuclease (CATH:FunFam 3.10.450.30:600), which contains 66 domains.

(domains) that compose the extended MSAs, when compared to the number of sequences that composed the STAMP structural MSAs. The distribution of the number of domain sequences in the structure-based MSAs generated by STAMP (Structural) was compared to that of the extended MSAs (Extended), for both CATH SCs and FunFams. The increase in the number of domain sequences with

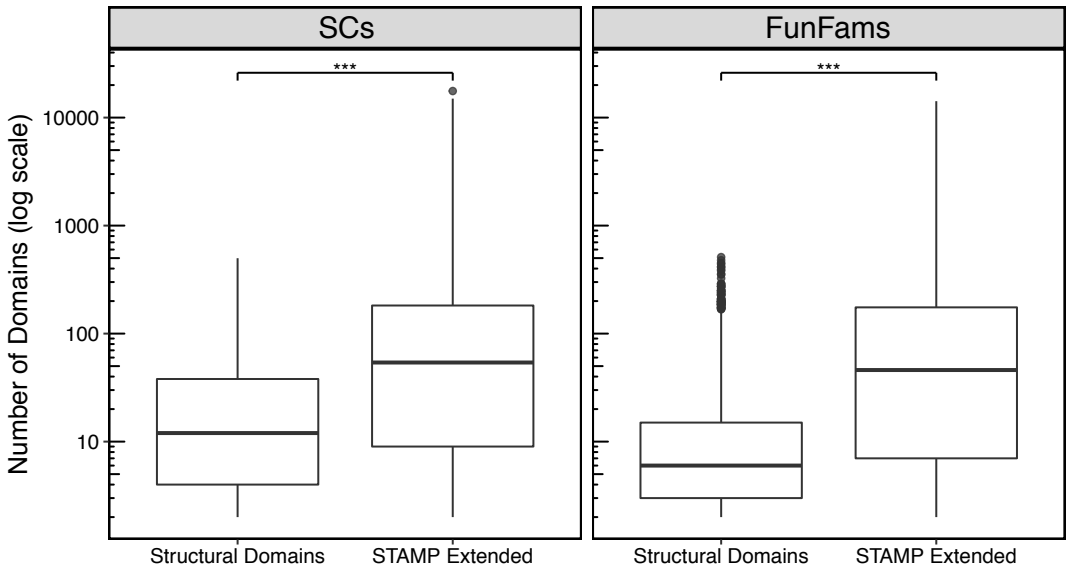


Figure 3.9: Box-plot showing the distribution of the number of domain sequences in the structure-based MSAs generated by STAMP (Structural) when compared to that of the extended MSAs (Extended), for both SCs and FunFams. The increase in the number of domains is statistically significant (p-value $< 1 \times 10^{-5}$).

the HMM-based protocol described in Section 3.3.2, is statistically significant for both SCs and FunFams (p-value < 1 x 10^{-5}). For CATH SCs a mean of 39 domain sequences constitutes the structure-based MSAs. This number increases to a mean of 312 sequences. Similarly, a mean of 16 domain sequences constitutes the MSAs generated for CATH FunFams, which is raised to a mean of 384 sequences. Although the number of domains in FunFams is on average lower than that of the SCs, the extended MSAs are on average larger for FunFams than for SCs. This likely results from the overall higher levels of sequence similarity observed among FunFam domain members (Figure 3.4), which results in a stronger signal in the profile HMMs generated. Remarkably, there is 25-fold and 32-fold overall increase, on average, of the number domains sequences that can be analysed for SCs and FunFams, respectively.

3.5 Conclusions

In order to perform an enriched analysis of genetic variants under a structure and evolutionary perspective (Chapter 4 and 5), structure-based MSAs were generated for CATH structural and functional families. A central point of focus in this Chapter was the careful assessment of the reliability of the structural MSAs generated by STAMP. This Chapter focused on improving the quality of the generated MSAs by: 1) developing methods that take advantage of the features of STAMP; and 2) exploring the power of HMMs for extending the alignments and annotating them with homologous protein sequences. Both these aspects are key for increasing the scope and coverage of the structure/sequence data currently available for analysis. The main conclusions of the work presented in this Chapter are:

- Selecting a seed domain from all-against-all domain STAMP scanning generates a good starting point for the alignment of multiple structures with STAMP.

- Further optimisation of the set of transformations obtained for the seed domain leads to an overall improvement of the *Sc* and RMSD alignment measures.

- Although STAMP produces overall reliable MSAs for CATH SCS and FunFams, the quality of the FunFam-based alignments is higher, which results from a higher structural similarity between the family domain members.

- STAMP-based MSAs were extended with related protein sequences using a new HMM-based protocol.

Chapter 4

Overall analysis of genetic variation

4.1 Summary

This Chapter focuses on the analysis of genetic variation in protein structural and functional families. Missense variants are analysed in the context of the structure-based MSAs generated for SCs and FunFams defined in CATH, as described in Chapter 3. Variants were collected and grouped according to their annotation and their attributes were compared. The distribution of nsSNPs regarding their structural locations was investigated. Missense variants were characterised across different structural sites and environments. Amino acid exchanges were investigated, and mutability matrices produced for germline, somatic and disease-associated variants. Genetic variation exchanges by annotation and the potential consequence were investigated for SC and FunFam domains. Analysis of exchanges in terms of amino acid change, physicochemical properties, as well as conservation was also performed.

Finally, analysis of prediction scores from SIFT and Polyphen-2 on the effects of the variants was performed.

4.2 Introduction

Missense mutations that gain clinical attention usually change the physicochemical properties of the amino acid residue sufficiently to affect the function of the gene product (Krawczak et al., 2000; Stone and Sidow, 2005), but the most severe mutations are likely to result in lethal phenotypes that cannot be inherited (Steward et al., 2003). At the same time, proteins are rather robust and can be quite tolerant to alterations in amino acid sequence (Poussu, 2005; Pajunen et al., 2007). In principle, an nsSNP can be deleterious either because it leads to disruption of a site that is directly involved in the function of a protein (e.g. a catalytic residue, a residue involved in ligand binding, or a residue that forms a critical interaction with another protein), or because it causes destabilisation of protein structure, leading to protein degradation, or the amino acid substitution abolishes protein function because of the loss of the structural framework that enabled the functionality of the protein in the first place. Pathogenic amino acid substitutions tend to have special characteristics that distinguish them from those nsSNPs that cause no phenotypic effect (neutral variations). In order to discriminate neutral variants from those causative of a disease phenotype, the prediction of the consequences of nsSNPs is a major research challenge, accompanying the rapid growth of genomic tools which have produced vast amounts of information about genetic variation within and among individuals (Ng and Henikoff, 2003; Karchin, 2009; Steward et al., 2003; Mooney, 2005)

4.2.1 Structural analysis of genetic variation

A general analysis of the effects of variation on the protein structures was performed by Wang and Moult, 2001. The effects of nsSNPs were grouped into protein stability, catalytic binding, allosteric regulation, ligand binding and post-translational modification. For protein stability, the factors investigated were the following: 1) loss of hydrogen bonds; 2) reduced hydrophobic interaction; 3) loss of a salt bridge; 4) introduction of a buried charge; 5) over-packing; 6) formation of an internal cavity (void space); 7) electrostatic repulsion; 8) causing the burial of a polar residue; 9) disruption of metal binding; 10) loss of disulphide bonds; 11) introduction of backbone strain; and 12) the destabilization of a protein interaction. The SNP dataset was separated into a disease subset, which contained variants known to be involved in disease, and a neutral subset. For the disease subset, the nsSNPs affected the protein stability of the protein more than the other factors. In contrast, for the neutral subset, the majority of the nsSNPs were found to have no effect on the structure. Furthermore, out of the six groups (protein stability, ligand binding, catalytic binding, allosteric regulation and post-translational modification), protein stability had the largest mutational effects (83%) for the disease subset. However, in the neutral set, the majority of nsSNPs were classified as having no effects and protein stability came second at 30%.

Following a similar approach, a global analysis of variation was also performed by Martin et al., 2002. The effect of mutations in the core region of the p53 protein was investigated. Seven effects of mutations on the p53 crystal structure were analysed: 1) mutations affecting a hydrogen bonding; 2) mutations to proline; 3) mutations

from glycine; 4) residue clashes; 5) mutation at the DNA binding site; 6) mutation at the zinc-binding site; and 7) mutation in conserved regions. An integrated pipeline to map single amino acid polymorphisms to structure was developed (Cavallo and Martin, 2005), which was later used in the development of SAAPdb (Hurst et al., 2009).

In order to understand the structural characteristics of SNPs, Stitziel et al., 2003, mapped a set of SNPs from OMIM and dbSNP to structures. Their objective was to introduce a new geometric classification for characterising disease SNPs. A set of SNPs from OMIM was classified as disease-causing while another set of SNPs from dbSNP was used as a control set, and both sets were mapped to protein structures. The majority of disease SNPs occurred in pocket or void regions (88%) and it was less likely that disease SNPs occur in convex regions. Additionally, disease SNPs tend to occur infrequently in completely buried regions. According to Stitziel et al., 2003, a mutation is less likely to be observed in the protein core since the core plays a critical role in protein stability; therefore, it is more likely that the mutation may be eliminated early in the stages of biogenesis.

Another study by Yue et al., 2006, analysed the structural location of SNPs. Homology modelling was used in this study to model human disease proteins before comparing them with SNPs from the UniProt database. The distribution of both disease and non-disease nsSNPs in protein 3D structures was analysed. A small set of 369 experimentally determined domains from the PDB were used, and 1,484 domains from 874 proteins were modelled, covering 6,352 mutations. The result showed a strong tendency for disease-associated variants to occur in the core region in patches (spatial clusters). Yue et al., 2006, suggested that disease SNPs forming

these patches could be involved in protein-protein interactions. In order to test this hypothesis, eight experimentally determined disease proteins were examined. In most of these experiments, disease mutations were observed to cluster at the interaction sites.

Gong and Blundell, 2010, recently catalogued structural and functional features of proteins that influence the substitution of amino acids. The motivation behind this study was to discover if the factors that restrain the substitution of amino acids in evolution also influence the occurrence of SNPs in coding regions. Although previous studies (Steward et al., 2003; Ng and Henikoff, 2003; Ferrer-Costa et al., 2005) look at the relationship between different factors (e.g., solvent accessible or conserved residue) with SNPs, Gong and Blundell, 2010, argued that most of these studies have not taken advantage of the rapidly growing information in structure and function.

4.3 Methods

4.3.1 Organising genetic variants according to their annotation

Genetic variants collected and organised in ProIntVar were separated into three main classes: 1) germline variants from a variety of sources, which include: dbSNP (Sherry et al., 2001); 1kGP (The 1000 Genomes Project Consortium, 2012); ExAC (Exome Aggregation Consortium (ExAC), Cambridge, MA, 2016); ESP (NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, 2016); HGMD-public (Stenson et al., 2014);

HapMap (The International HapMap Consortium, 2007); and PhenCode (Giardine et al., 2007); 2) somatic variants from COSMIC (Forbes et al., 2015); and 3) disease-associated (a subset of germline variants), obtained from: Humsavar (Wu et al., 2006; Famiglietti et al., 2014); OMIM (Amberger et al., 2015); and ClinVar (Landrum et al., 2016). Section 2.3.13 overviews the source and set of annotations collected for each variant. The variation types covered are: missense (non-synonymous single nucleotide polymorphisms (nsSNPs)), frameshift-variant, stop-gained, start-lost, inframe-deletion, inframe-insertion, splice-region-variant. Synonymous SNPs were not considered in this study, since the mutational outcome results in the same amino acid.

The consequences of genetic variation were investigated using prediction scores and categorical classification from Polyphen-2 (Adzhubei et al., 2010) and SIFT (Kumar et al., 2009). Only these two methods were used, since prediction scores were readily available for each variant from the Ensembl database (Chen et al., 2010; Flicek et al., 2013; Cunningham et al., 2015), through the Ensembl REST API (Rios et al., 2010; Yates et al., 2014), as well as from the UniProt Variation endpoint (The UniProt Consortium, 2015). Qualitative categorical prediction states are defined from the prediction scores (P) provided by SIFT as: deleterious ($P \leq 0.05$); or tolerated ($P > 0.05$). Similarly, categorical states generated for Polyphen-2 (HumVar) are classified as: benign ($P < 0.446$); probably damaging ($0.446 \leq P < 0.909$); or possibly damaging ($P \geq 0.909$).

4.3.2 Characterising genetic variants across structural regions and environments

Genetic variants were mapped to protein 3D structure as described in Section 2.3.3 and Section 2.3.12. Protein domains were grouped into SCs and FunFams in CATH (Sillitoe et al., 2015) and MSAs generated by the structural alignment of the domains (Section 3.3.1). MSAs were further extended with similar protein sequences (Section 3.3.2). The analysis of genetic variants focused on three protein sequence/structure datasets including: protein domains classified in CATH that are mapped to protein chains; protein domains that are mapped to protein chains in the PDB and are not covered by CATH; and finally protein sequence domains that were found to extend the STAMP structural alignments (Section 2.3.13). Mutated residues and their position in 3D space was investigated in the context of the SC/FunFam protein families.

Figure 4.1 illustrates the process of mapping genetic variants and structure-annotated residues onto the STAMP extended MSAs. Residues were annotated according to their location within the protein structural regions (environments) as described in Section 2.3.7. The occurrence of genetic variants was compared across secondary structure elements (SSEs): *α-helix*; *β-sheet/strand*; and *Turns or coil*; across structural spatial environments: solvent inaccessible (*Core*), partially exposed to solvent (*Part. Exposed*), accessible to the solvent (*Surface*), and interaction interface (*Interface*): interaction with domains (*Inter. Domain*), interaction with ligands (*Inter. Ligand*, ligand definition is provided in Section 2.3.5), and interaction with other protein residues which are not part of CATH domains (*Inter. Protein*);

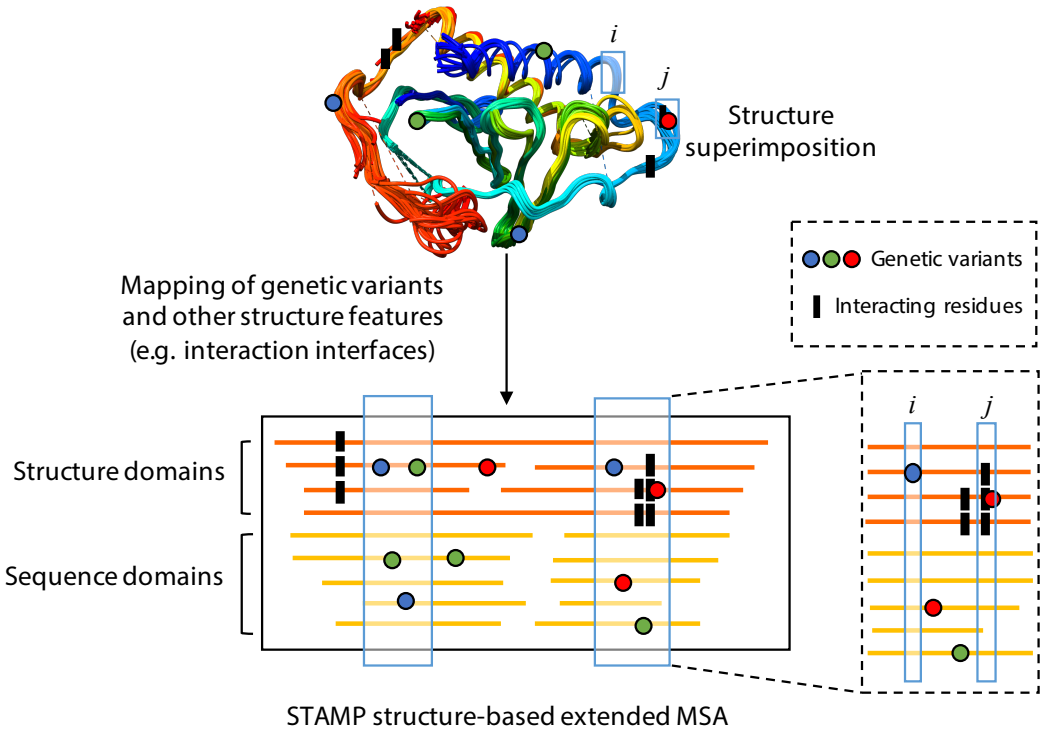


Figure 4.1: Mapping of genetic variants and structure-annotated residues in structure-based MSAs. MSAs are generated by STAMP and extended with similar sequences of proteins with unknown structure as described in Section 3.3. Structurally conserved columns/positions in the structure-based domain MSA are defined by STAMP. Conserved regions as well as individual MSA columns, i and j, are highlighted. The mapping of genetic variants and structure-annotated residues is illustrated for both MSA and the structure superimposition.

as well as across STAMP-defined structurally equivalent (*Conserved*) regions (or positions) (SCRs), and *Variable* regions, as described in Section 3.3.1.

4.3.3 Mapping and analysis of variants in the MSAs

As shown in Figure 4.1, multiple genetic variants and structure features are mapped onto aligned positions (or columns) in the MSA. The characterisation of genetic variants that map onto particular structure domains was performed considering the structural annotations of that particular domain alone. This is particularly important for interaction interfaces, which were treated differently so that only domains

which are found to interact are annotated as such. As an example, the majority of domain residues that are aligned in column i are classified to be part of an α-helix, nevertheless, a variant position is classified as turns/coil if the secondary structure state of the domain to which the variant is mapped to is structurally annotated with that state. In addition to secondary structure classification, variant positions in column i are also annotated with regard to the spatial location, accessibility to solvent, as well as to the STAMP conservation status of the domains to which the variation maps to.

STAMP-defined structurally conserved regions (SCRs) (columns), as well as variable regions were analysed as structural environments. SCRs were used to infer conserved regions in the similar protein sequences found to extend the structure-based MSAs (Figure 4.1). Characterisation of genetic variants was conducted so that every unique variant was only considered once in the analysis (Section 2.3.13). This prevents redundancy arising from multiple structure-sequence mapping being propagated or considered multiple times.

Variants were compared across the structural regions and environments using the odds ratio (OR) measure described in David et al., 2012. The preference for a variant to be in region i rather than region j was calculated as:

$$OR_{ij} = \frac{p_i/(1-p_i)}{p_j/(1-p_j)} \tag{4.1}$$

where the probability of observing variants in region i is $p_i = \frac{n_i}{N_i}$ (n_i corresponds to the number of observed variants, whereas N_i corresponds to the total number of available residues in region i). To compare two structural environments, log OR

scores are calculated as the ratio between the odds (relative likelihood) $p_i/(1-p_i)$ and $p_j/(1-p_j)$. When the total number of variants and residues in analysis are identified within multiple regions (e.g. *α-helix*, *β-sheet*, and *Turns or coil*), variants and residues in a region i (e.g. *α-helix*) are compared to all other variants/residues that compose the remaining regions, aside from the region under analysis.

The mutability score for each amino acid residue was calculated by taking the total number of mutations observed in the dataset divided by its frequency of occurrence in the human proteome. Background relative amino acid frequencies for the human proteome were obtained from de Beer et al., 2013, and are summarised in Table 2.5. Inspired by the Point Accepted Mutation (PAM) matrices generated by Dayhoff and Schwartz, 1978, as well as the Block Substitution Matrices (BLOSUM) generated by Henikoff and Henikoff, 1992, log-odds mutability matrices were calculated for the genetic variants as:

$$R_{ij} = \left(\frac{1}{\lambda}\right) \log\left(\frac{P(ij)}{P(i)P(j)}\right) \tag{4.2}$$

where $P(ij)$ is the probability of two amino acids i and j replacing each other in a similar sequence, and $P(i)$ and $P(j)$ are the background probabilities of finding the amino acids i and j in any protein sequence. An arbitrary scaling factor of $\lambda = 2$ was set such that the matrix contains easily comparable values.

The Shannon's entropy (H) (Shannon, 1948) conservation measure was calculated for each alignment column position as:

$$H = -\sum_{i}^{20} p_C(i) log(p_C(i)) \tag{4.3}$$

where p_C is the distribution of a set of 20 amino acids in column C. H is subsequently calculated using the frequency of each amino acid i observed in column C as $p_C(i)$, and is the smallest for a column with complete conservation ($H = 0.0$).

4.3.4 Analysis of variation exchanges for amino acids and their physicochemical properties

Figure 4.2 shows a summary Venn diagram with the physicochemical properties of amino acids as defined by Taylor, 1986, and Livingstone and Barton, 1993. Two alphabets that group amino acids according to physicochemical properties were explored. In the first alphabet, 'Chemical_A', amino acids were divided into the following six groups according to their physicochemical properties as: polar (Gln, Asn, His, Ser, Thr, Tyr, Cys, Met and Trp); hydrophobic (Ala, Ile, Leu, Phe, Val); basic (Arg and Lys); acidic (Asp and Glu); proline (pro); and glycine (Gly). Pro and

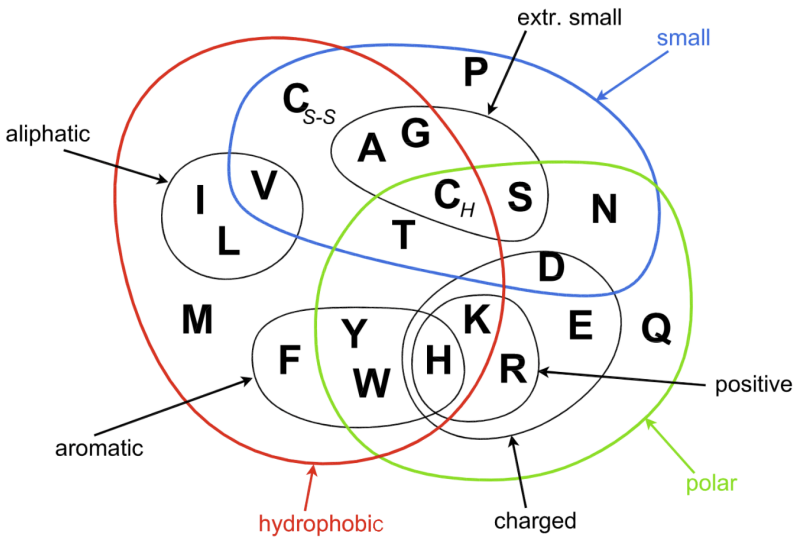


Figure 4.2: Summary of the physicochemical properties of amino acids. Amino acids are grouped according to overlapping properties which include: size, charge, polarity and hydrophobicity. Figure adapted from Taylor, 1986, and Livingstone and Barton, 1993.

Gly amino acids were analysed as separated classes in this alphabet as their particular importance in disease-association has been reported (Martin et al., 2002). The second alphabet, 'Chemical_B', grouped amino acids as: aliphatic (Leu, Ala, Gly, Val, Ile, Pro); acidic (Glu, Asp); basic (Arg, Lys, His); small hydroxy (Ser, Thr); aromatic (Phe, Tyr, Trp); amide (Gln, Asn); and sulphur (Met, Cys). Although other amino acid classifications could have been used (e.g. according to size), these simple classifications were used as they allow to classify all amino acids without resulting in amino acid overlap between classes.

Standard amino acid properties summarised in Table 2.5, were used in the various analyses performed in this work. Amino acid properties include: monoisotopic atomic mass (Knapp, 1996); average volume of buried residues, calculated from the surface area of the side chain (Zamyatnin, 1972); hydrophobicity (Fauchère et al., 1988); total accessible surface area (Ahmad et al., 2004); and frequency of occurrence (abundance) in the human proteome (de Beer et al., 2013).

4.4 Results and discussion

4.4.1 Mapping genetic variation to SCs and FunFam domains

Genetic variants were grouped by annotation into three main classes: germline variants, somatic variants and disease-associated variants. Table 4.1 summarises the number of SCs and FunFams for which genetic variants could be obtained, as well as the mean number of variants per variation group. Genetic variation was mapped

to domain members of 663 CATH SCs and 1,231 FunFams, from a total of 2,371 SCs and 5,777 FunFams, respectively. The resulting mean number of variants is 210 and 124, for SCs and FunFams, respectively. Taking all unique genetic variants together results in a total of 139,516 for SCs and 152,926 for FunFams. The highest number of unique variants mapped across SC/FunFam families is 7,511. Figure 4.3 shows a positive Pearson correlation ($r = 0.41$ and $r = 0.20$, for SCs and FunFams, respectively) between the number of human missense variants mapped to domain members of SCs and FunFams and the number of domains per SC/FunFam family. This result indicates that a higher number of domain residues per family increases the likelihood of observing a higher number of variant positions. Some protein families seem to be particularly enriched for variants, whereas others show only a small number of variants mapped to them (Figure 4.3). These outlier protein families result from the fact that not all SC/FunFam families are composed of human proteins, for which variation is obtained. The total number of domains per family also affects the correlation, despite the fact that only uniquely mapped variants are considered and the fact that, on average, only 2% of the domains per family have variants mapped to them.

Table 4.1: Overview of SCs and FunFams for which genetic variation could be mapped to their domain members. Genetic variants were grouped as germline variants, somatic mutations and disease-associated variants. The mean ($\bar{x}$) [minimum; median; maximum] and the total number of genetic variants that were mapped to domains in SCs and FunFams is shown.

	SCs			**FunFams**		
Class of variants	**Numb. of SCs**	**Variants per SC ($\bar{x}$)**	**Total numb. of variants**	**Numb. of FunFams**	**Variants per FunFam ($\bar{x}$)**	**Total numb. of variants**
All[a]	663	210 [1; 114; 7,511]	139,516	1,231	124 [1; 76; 7,511]	152,926
Germline[a]	662	133 [2; 75; 2,784]	87,769	1,230	78 [1; 49; 2,784]	96,096
Somatic[a]	658	66 [1; 27; 3,499]	43,325	1,217	39 [1; 20; 3,499]	47,756
Disease[a]	336	26 [1; 7; 1,228]	8,422	495	18 [1; 5; 1,228]	9,074
All[b]	728	408 [1; 127; 57,220]	296,887	1,822	365 [1; 117; 22,870]	664,175
Germline[b]	728	284 [1; 90; 39,580]	206,358	1,814	256 [1; 84; 16,020]	464,508
Somatic[b]	703	117 [1; 31; 17,340]	82,470	1,744	103 [1; 29; 6,678]	179,827
Disease[b]	388	21 [1; 6; 535]	8,059	901	22 [1; 7; 537]	19,840
All[c]	662	187 [1; 101; 5,680]	123,729	1,231	111 [1; 70; 5,693]	137,060
Germline[c]	662	120 [2; 69; 2,185]	79,157	1,230	71 [1; 45; 2,193]	87,268
Somatic[c]	661	56 [1; 24; 2,284]	37,050	1,215	34 [1; 18; 2,285]	41,505
Disease[c]	325	23 [1; 6; 1,211]	7,522	477	17 [1; 5; 1,215]	8,287

[a]All variation types mapped onto protein structure domains; [b]All variant types mapped onto similar protein sequences with unknown structure; [c]Final subset (as described in the text) of human missense mutations mapped onto protein structure domains.

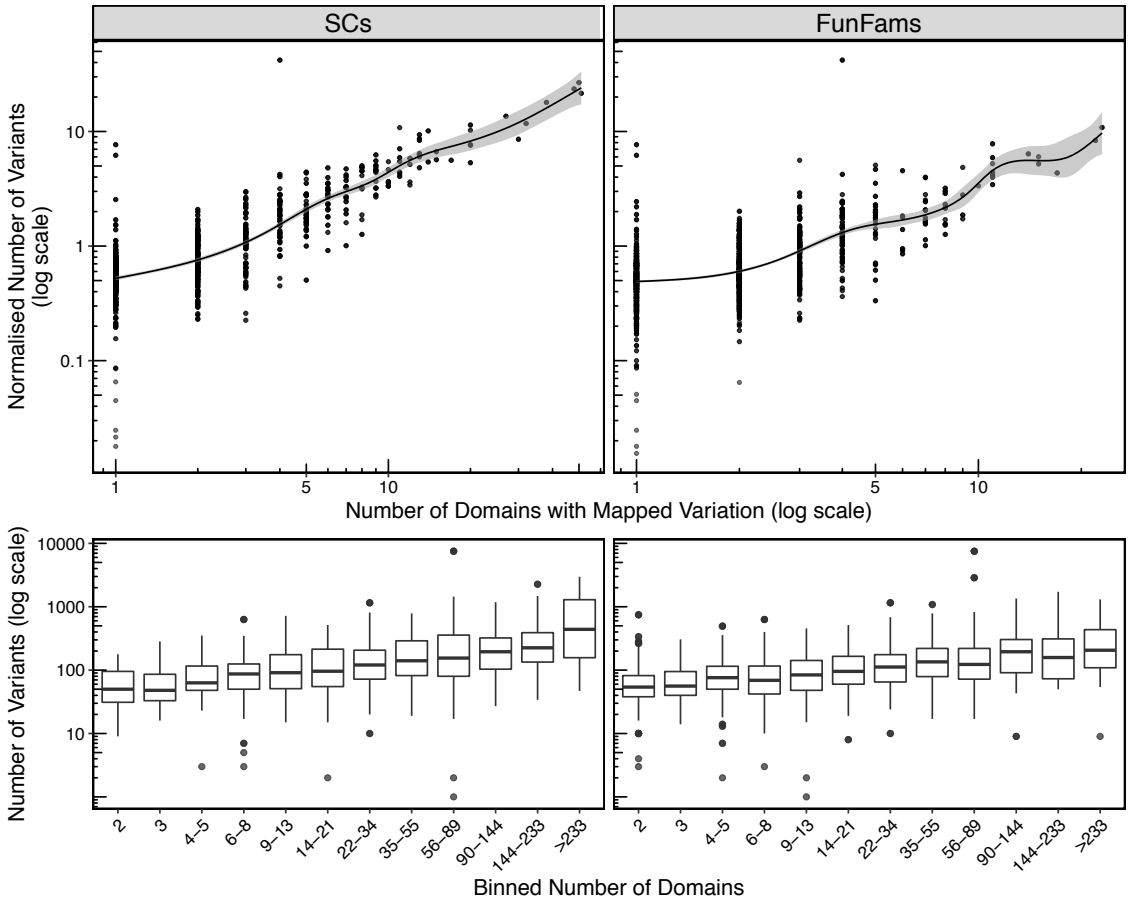


Figure 4.3: Correlation between the number of genetic variants mapped to SCs and FunFams domains and the number of domains per SC/FunFam for which variation could be obtained. The binned total number of domains in each SC/FunFam was also compared to the number of mapped variants.

The breakdown of the number of variants in SCs and FunFams by the class of variation is also provided in Table 4.1. The mean number of variants per class drops to 133 and 78 for germline variants, 66 and 39 for somatic variants, and 26 and 18 for disease-associated variants, for SCs and FunFams, respectively. Germline variants account for 63% of the total variants analysed in both SCs and FunFams. Somatic variants account for 31%, whereas known disease-associated variants account for only 6%. CATH Superfamily 2.60.40.720 (SC 2 and FunFam 7) accounts for the highest number of somatic and disease-associated variants, mapped to the structure domains (Figure 4.3). This Superfamily comprises the Cellular tumour antigen p53,

which is the most frequent target for mutation in human cancer (Freed-Pastor and Prives, 2012), among other proteins. p53 corresponds to the most widely studied protein (Prives and Hall, 1999), so observational bias is expected for this protein family.

Table 4.1 also summarises the number of variants that were mapped to similar protein sequences in the extended MSAs. The number of SCs and FunFams for which variants could be mapped increased from 663 to 728 and 1,231 to 1,822, resulting in a mean number of variants of 408 and 365, for SCs and FunFams, respectively. This result was expected since the number of sequences found in the MSAs was extended by the HMM-based protocol (Section 3.4.3). The increase in the availability of genetic variants introduced by extending the MSAs enables a much higher number of genetic variants to be analysed in the context of the SCs and FunFams. In fact, an additional 296,887 and 664,175 human genetic variants were identified for SCs and FunFams, respectively. A higher proportion of these additional variants corresponds to germline variants (69% compared to 63% in structure domains), with 28% somatic and 3% disease-associated variants. Although preliminary analysis of the STAMP MSAs indicate that STAMP-defined structurally conserved positions/regions (Section 3.4.2) could be used to infer structure domain features onto similar protein sequences with unknown structure, additional analysis need to be performed in order to extrapolate information between the structures and sequences within measurable levels of confidence. Therefore, only variants that were mapped to protein structure domains in the SCs and FunFams were considered in the remaining analysis performed here.

From the human genetic variants mapped onto SCs and FunFams, 89% are of

the missense type (nsSNPs). The remaining variants correspond to a combination of frameshift variants, stop-gained, stop-lost, in-frame deletion, in-frame insertion, and splice-region variants. Table 4.2 overviews the total number of frameshift, stop-gained and stop-lost mutations mapped onto structure domains in SCs and FunFams. Although these variation types are among those that are regarded as potentially producing more dramatic protein phenotypes (Stankiewicz and Lupski, 2010; Gonzaga-Jauregui et al., 2012), only between 3 to 6% of these variants are currently annotated as disease-associated. These variant types were not included in the remaining analysis, since only 'from' and 'to' exchanges where a single amino acid is mutated were considered. Nevertheless, these variants are kept in ProIntVar for analysis. In fact, some examples of frameshift and stop-gaining mutations are further explored in Section 5.4.8.

The final subset of missense variants from human mapped onto protein structure domains corresponds to a total of 123,729 variants, with a mean of 187 variants per SC across 662 SCs (Table 4.1). This corresponds to a total of 137,060 unique variants and 1,231 FunFam protein families, and a mean of 111 variants per FunFam (Table

Table 4.2: Total number of frameshift and stop-gained/lost variants in SCs and FunFams across the three main classes of variants.

	SCs			**FunFams**		
Class of variants	**Frameshift**	**Stop gained**	**Stop lost**	**Frameshift**	**Stop gained**	**Stop lost**
All	5,738	6,004	29	6,200	6,559	27
Germline	3,293	2,455	25	3,732	2,770	23
Somatic	2,177	3,162	3	2,197	3,363	3
Disease	268	387	1	271	426	1

4.1). The relative proportion of variants by class is identical to the initial dataset with 64% germline nsSNPs, 30% somatic mutations, and 6% disease-associated variants (a part of germline or somatic variants), for both SCs and FunFams. Table 4.3 highlights the top-10 most mutated SC/FunFam families. The top protein families include protein domains from all three main classes in CATH, mainly α-helix (Class 1), mainly β-sheet (Class 2), and mixed secondary structure content (Class 3). Five of the top-10 families comprise domains that interact with nucleic acids (e.g. Winged-helix DNA binding domain family).

Figure 4.4 shows an example of a STAMP structure-based MSA and structure superimposition where mapping of variants is performed with the aid of ProIntVar. The example is for the CATH Superfamily 1.10.540.10 (Butyryl-CoA dehydrogenase, subunit A, domain 1, SC ID 1 and FunFam ID 15764). This family is composed of Medium-chain-specific acyl-CoA dehydrogenases (EC 1.3.99.3).

Table 4.3: Top-ranking SC and FunFam protein families for which genetic variants could be mapped to its domain members.

Rank	SPF ID	SC ID	FunFam ID	Description[a]
1	2.40.50.100	3	50450	Sulfate ABC transporter
2	1.10.10.10	24	266362	Winged-helix DNA binding domain
3	3.10.100.10	3	11954	Mannose-binding protein
4	1.10.10.10	9	123911	Winged-helix DNA binding domain
5	3.30.160.70	1	2113	Protein-cysteine DNA methyltransferase
6	3.30.40.10	12	40330	Zinc/RING finger domain (C3HC4)
7	2.30.18.10	1	280	Transcription factor IIA β-barrel
8	1.10.287.1010	1	31	Ataxin 3
9	2.10.25.10	4	45504	Laminin
10	1.10.10.60	7	124109	Homeodomain-like

[a]CATH Superfamily (SPF) description or representative domain name.

10 20 30 40 50 60 70 80 90 100
2uxw_A_02 1 - P - - S - - - - V L N E E Q - T Q F L K E L V E P V S R F F E E V N - - D P A K N D A L E M V - E - E - T T W Q G L K E L G A F G L Q V P S E L G G V G - - - L - - C N T Q Y 70
3p4t_B_01 1 - I - - - - W T - - - - - T A E R E A L R K T V R A F A E R E V L P H A H E W E R A G E I - P - R - E L H R K A A E L G L L G A G F P E D A G G S G - - - G - - D G A D P 67
3r7k_D_01 1 - A - - - - W T - - - - - T P E R R A L S Q M A R S F V E R E I A P K L A E W E H V G E I - P - R - D L H L N A A E V G L L G I G F P E E V G G S G - - - G - - N A I D S 67
3owa_B_01 1 A V - - - K G G S F L V D E I T I D Q V F - T - - P E D F - - - - - S S E H K X I A K T T E D F I V N E V L P E L E Y L E Q - H E F D R S V - R L L K E A G E L G L L G A D V P E E Y G G I G - - - L - - D K V S S 88
2z1q_B_01 1 K K L W Q K G G G W L L E V P E - - R V Y - T - - P E D F - - - - - D E S V K E I A R T T R T F V E R E V L P L L E R M E H - G E L E L N V - P L M R K A G E L G L L A I D V P E E Y G G L D - - - L - - P K V I S 89
1ukw_A_01 1 - - - - - - - - - - - - - - - - - - - I D - F - - - - S L - - - - - T E E Q R Q L Q A L A R R F A K E V I L P V A Q E Y D E K E E V - P - W - P V I E K L H E V G L L N A I I P E E Y G G M G - - - L - - K M L D E 69
1udy_B_01 1 - - - - - - - - - - - - - - - - - - - G F S - F - - - - E L - - - - - T E Q Q K E F Q A T A R K F A R E E I I P V A A E Y D R T G E Y - P - V - P L L K R A W E L G L M N T H I P E S F G G L G - - - L - - G I I D S 70
1egc_D_01 1 - - - - - - - - - - - - - - - - - - - F S - F - - - - E F - - - - - T E Q Q K E F Q A T A R K F A R E E I I P V A A E Y D K T G E Y - P - V - P L I R R A W E L G L M N T H I P E N C G G L G - - - L - - G T F D A 69
2dvl_B_01 1 - L - - - - - T Q E Q R L V L D A V R R V A R E V L Y P L A P E Y D R K A E Y - P - W - P Q L K A L A E L G L L G M T T P E E W G G V G - - - L - - D S V T W 65
3nf4_B_01 1 - - - - - - - - - - - - - - - - - - D - R - - - - - - - L L - - - - - P S Q E A A E L I E L T R E I A D K V L D P I V D R H E K D E T Y - P - E - G V F E Q L G A A G L L S L P Q P E E W G G G G - - - Q - - P Y E V Y 69
4iv6_B_01 1 - A L - - - - - T A E E E T I V K T V H D F V E K Q V K P V V R E L E H A N T Y - P - E - E L I E T M K E I G I F G L A I P E P Y G F G A - - - V - - S M P C Y 66
2cx9_C_01 1 - F - - - - E E - - - - - G A E E R Q V L G P F R E F L K A E V A P G A A E R D R T G A F - P - W - D L V R K L A E F G V F G A L V P E A Y G G A G - - - L - - S T R L F 67
1buc_B_01 1 - M D - F - - - - N L - - - - - T D I Q Q D F L K L A H D F G E K K L A P T V T E R D H K G I Y - D - K - E L I D E L L S L G I T G A Y F E E K Y G G S G D D G G - - D V L S Y 72
1rx0_D_01 1 - - - - - - - - - - T S C I - - - D P S M - - - - - - G L - - - - - N E E Q K E F Q K V A F D F A A R E M A P N M A E W D Q K E L F - P - V - D V M R K A A Q L G F G G V Y I Q T D V G G S G - - - L - - S R L D T 74
2jif_A_01 1 - A P L - - Q - T F - - - - - T D E E M M I K S S V K K F A Q E Q I A P L V S T M D E N S K M - E - K - S V I Q G L F Q Q G L M G I E V D P E Y G G T G - - - A - - S F L S T 70
2pg0_B_01 1 - R - - - - Y L - - - - - R E E H H M F R A A F R K F L E K E A Y P H Y N D W E K R G I I - P - R - S F W A K M G E N G F L C P W V D E K Y G G L N - - - A - - D F A Y S 67
3pfd_A_01 1 - E H I A L R E A I R A L A E K E I A P Y A A E V D E K A R F - P - E - E A L A A L N S S G F S A I H V P E E Y G G Q G - - - A - - D S V A T 62
1jqi_A_01 1 - Q S V - - - - E L - - - - - P E T H Q M L R Q T C R D F A E K E L V P I A A Q L D K E H L F - P - T - S Q V K K M G E L G L L A M D V P E E L S G A G - - - L - - D Y L A Y 69
3mpi_A_01 1 - F - - - - N L - - - - - S K E L Q M L Q K E V R N F V N K K I V P F A D Q W D N E N H F - P - Y E E A V R P M G E L G F F G T V I P E E Y G G E G - - - M D Q G W L A A 70
1ivh_C_01 1 - A I - N - - - - G L - - - - - S E E Q R Q L R Q T M A K F L Q E H L A P K A Q E I D R S N E F K N L R - E F W K Q L G N L G V L G I T A P V Q Y G G S G - - - L - - G Y L E H 71
1sir_A_01 1 - - - - - - - - - - E F D W Q D P L - - - V - L E E - - Q L - - - - - T T D E I L I R D T F R T Y C Q E R L M P R I L L A N R N E V F - H - R - E I I S E M G E L G V L G P T I K - G Y G C A G - - - V - - S S V A Y 77
110 120 130 140 150
2uxw_A_02 71 A R L V E E I V G M - H D - L G V G I T L G A H Q S I G F K G T L L F G T K A Q K E K Y L P K L A S G E T V 121
3p4t_B_01 68 V V I C E E M H Y A G S P G G V Y A S L F T C G - I A V P H M I A S G D Q R L I D T Y V R P T L R G E K I 119
3r7k_D_01 68 A L V T E A I L A A G G S T G V C A A L F T H G - I A L P H I A A N G S D A L I E R Y V R P T L A G K M I 119
3owa_B_01 89 A L I A E K F S R - - A - G G F A I T H G A H V G I G S L P I V L F G N E E Q K K K Y L P L L A T G E K L 138
2z1q_B_01 90 T V V A E E L S G - - S - G G F S V T Y G A H T S I G T L P L V Y F G T E E Q K R K Y L P K L A S G E W I 139
1ukw_A_01 70 V I V G E E L A Y - A C - M G I Y T I P M A S D - L G I T P V L L A G T E E Q K E R F L R P L T E K - - - 116
1udy_B_01 71 C L I T E E L A Y - G C - T G V Q T A I E A N T - L G Q V P L I I G G N Y Q Q Q K K Y L G R M T E E P L M 120
1egc_D_01 70 C L I S E E L A Y - G C - T G V Q T A I E G N S - L G Q M P I I I A G N D Q Q K K Y L G R M T E E P L M 119
2dvl_B_01 66 A L A L E E L A A - A D - P S V A V I V S V T S G L P Q Y M L L R F G S E A Q K R R Y L V P L A R G E W I 116
3nf4_B_01 70 L Q V L E E I A A - R W - A S V A V A V S V H S - L S S H P L L V F G T E E Q K K R W L P G M L S G E Q I 119
4iv6_B_01 67 V Q V A E E L A R - G W - M S L A G A M G G H T - V V S K L L L L F G T E E Q K Q K Y L P R M A T G E L R 116
2cx9_C_01 68 A R M V E A I A Y - Y D - G A L A L T V A S H N S L A T G H I L L A G S E A Q K E A F L P K L A S G E A L 118
1buc_B_01 73 I L A V E E L A K - Y D - A G V A I T L S A T V S L C A N P I W Q F G T E A Q K E K F L V P L V E G T K L 123
1rx0_D_01 75 S V I F E A L A T - G C - T S T T A Y I S I H N - M C A W M I D S F G N E E Q R H K F C P P L C T M E K F 124
2jif_A_01 71 V L V I E E L A K - V D - A S V A V F C E I Q N T L I N T L I R K H G T E E Q K A T Y L P Q L T T E - - - 118
2pg0_B_01 68 V V I N E E L E K - V G - S S L - V G I G L H N D I V T P Y I A S Y G T E E Q K Q K W L P K C V T G E L I 117
3pfd_A_01 63 C I V I E E V A R - V D - C S A S L I P A V N K - L G T M G L I L R G S E E L K K Q V L P A V A S G E A M 112
1jqi_A_01 70 S I A L E E I S R - G C - A S T G V I M S V N N S L Y L G P I L K F G S S Q Q K Q Q W I T P F T N G D K I 120
3mpi_A_01 71 M I V T E E I A R - G S - S A L R V Q L N M E V L G C A Y T I L T Y G S E A L K K K Y V P K L S S A E F L 121
1ivh_C_01 72 V L V M E E I S R - A S - G A V G L S Y G A H S N L C I N Q L V R N G N E A Q K E K Y L P K L I S G E Y I 122
1sir_A_01 78 G L L A R E L E R - V D - S G Y R S A M S V Q S S L V M H P I Y A Y G S E E Q R Q K Y L P Q L A K G E L L 128
2uxw_A_02
3p4t_B_01
3r7k_D_01
3owa_B_01
2z1q_B_01
1ukw_A_01
1udy_B_01
1egc_D_01
2dvl_B_01
3nf4_B_01
4iv6_B_01
2cx9_C_01
1buc_B_01
1rx0_D_01
2jif_A_01
2pg0_B_01
3pfd_A_01
1jqi_A_01
3mpi_A_01
1ivh_C_01
1sir_A_01

Figure 4.4: Example of a STAMP structure-based MSA, dendrogram and structure superimposition highlighting variant positions. The example is for Medium-chain-specific acyl-CoA dehydrogenase (CATH Superfamily 1.10.540.10, FunFam ID 15764). Germline (orange), somatic (red) and disease-associated (dark red) variants are highlighted in both the MSA and domain superimposition. Structure superimposition is provided for the domains highlighted in bold text. MSA colouring is based on BLOSUM62 scores as implemented in Jalview (Waterhouse et al., 2009). Dendrogram was calculated by Jalview using Neighbour Joining PID.

4.4.2 Analysis of genetic variation across different structural environments

In section 4.3.2 various structural sites/regions were defined for the analysis of variants. These include secondary structure elements (SSEs): α-helix, β-sheet (or strand), and turns/coils; residue spatial subsets according to the solvent accessibility and interaction: core, partially exposed surface, surface, and interface; as well as structurally equivalent (conserved) or variable residues, as defined by STAMP (Section 3.3.1).

Table 4.4 summarises the number of amino acid residues that compose each of these structure regions. Residue counting is only cumulative within each of the subsets defined, i.e. within secondary structure elements, spatial location, and STAMP structure conservation. Additionally, residues that are located at interaction interfaces are organised according to three types of interaction: interaction with domains (domain-domain interaction); interaction with ligands (domain-ligand); and interaction with other protein residue partners which are not part of domain themselves (domain-protein). The mean number of residues and distributions across the different regions/environments is identical for both SCs and FunFams. Core, partially exposed and surface spatial locations account for 15%, 14% and 40% of the residues

Table 4.4: Residue composition of SC and FunFam domains across different structural environments. The mean ($\bar{x}$) [minimum; median; maximum] and the total number of residues mapped to SCs and FunFams is provided.

	SCs		**FunFams**	
Region	**Residues per SC ($\bar{x}$)**	**Total residues**	**Residues per FunFam ($\bar{x}$)**	**Total residues**
Conserved	108 [14; 88; 452]	70,747	133 [28; 114; 457]	163,880
Variable	137 [0; 100; 697]	90,686	39 [0; 18; 551]	48,234
Core	36 [0; 26; 212]	24,064	33 [0; 24; 189]	40,906
Exposed	34 [0; 27; 166]	22,359	30 [0; 24; 170]	36,720
Surface	98 [7; 81; 442]	64,765	69 [8; 59; 339]	84,972
Interface	76 [0; 63; 411]	50,245	40 [0; 29; 411]	49,516
Inter. Domain[a]	48 [0; 40; 287]	31,745	27 [0; 21; 289]	32,868
Inter. Ligand[a]	12 [0; 6; 110]	7,936	6 [0; 3; 108]	7,726
Inter. Protein[a]	16 [0; 7; 190]	10,564	7 [0; 1; 190]	8,922
α-helix	88 [0; 72; 344]	58,023	62 [0; 50; 572]	75,954
β-sheet/strand	41 [0; 36; 263]	27,022	37 [0; 38; 226]	45,438
Turns or coil	115 [1; 92; 533]	76,388	77 [1; 58; 453]	90,722

[a]Interface residues which result from the interaction with: domains, ligands, and other proteins (not domain); are a subset of the interface.

in SCs. For FunFams, 19% core, 17% partially exposed and 40% surface residues are observed. A mean of 76 and 40 residues in SCs and FunFams, respectively, participate in interactions. Although the number of alignment positions (columns) where interaction residues are found is higher for SCs than for FunFams, the proportion across different types of interaction is relatively similar. Domain-domain interactions account for 63% and 67% of the interface residues for SCs and FunFams, respectively. Domain-ligand and domain-protein interactions account for 15% and 18-21% of the residues, respectively. Regarding secondary structure elements, 35-36% residues in

SCs and FunFams are annotated as α-helix, whereas 17-21% and 44-47% are annotated as β-sheet and turns/coils, respectively. These SSE proportions broadly agree with previous analysis (Sreerama et al., 1999). A bigger difference is observed for the mean number of residues in STAMP conserved/variable regions and in interaction interfaces. As discussed in Chapter 3.4.1, the number of FunFams is triple the number of SCs in CATH. This means that the number of domains clustered into SCs is higher than those classified into FunFams. As a result of this, longer alignments are produced for SCs due to a higher number of domains and resulting insertion of gaps, when compared to FunFam MSAs (Figure 3.4 A). Accordingly, the mean number of residues in STAMP variable regions is 137 for SC and only 39 for FunFam. In contrast, the number of conserved residues is 133 for SCs and 108 for FunFams. This results from the fact that less structural diversity is observed for FunFams than for SCs, as a result of grouping together domains on the basis of functional similarity in addition to structural similarity. This leads to a higher number of structurally conserved positions being identified for FunFams than for SCs (Figure 3.6).

Figure 4.5 summarises the distribution of unique genetic variants mapped to domains within SCs and FunFams across the various structural environments. The distribution profile is overall similar for variants in all three classes, as well as between SCs and FunFams. The number of residues that compose each structure environment and the number of variants that are mapped to them is positively correlated for both SCs and FunFams ($r = 0.93$ and $r = 0.95$, respectively). This result is expected since a bigger region is expected to accommodate a higher number of mapped variants.

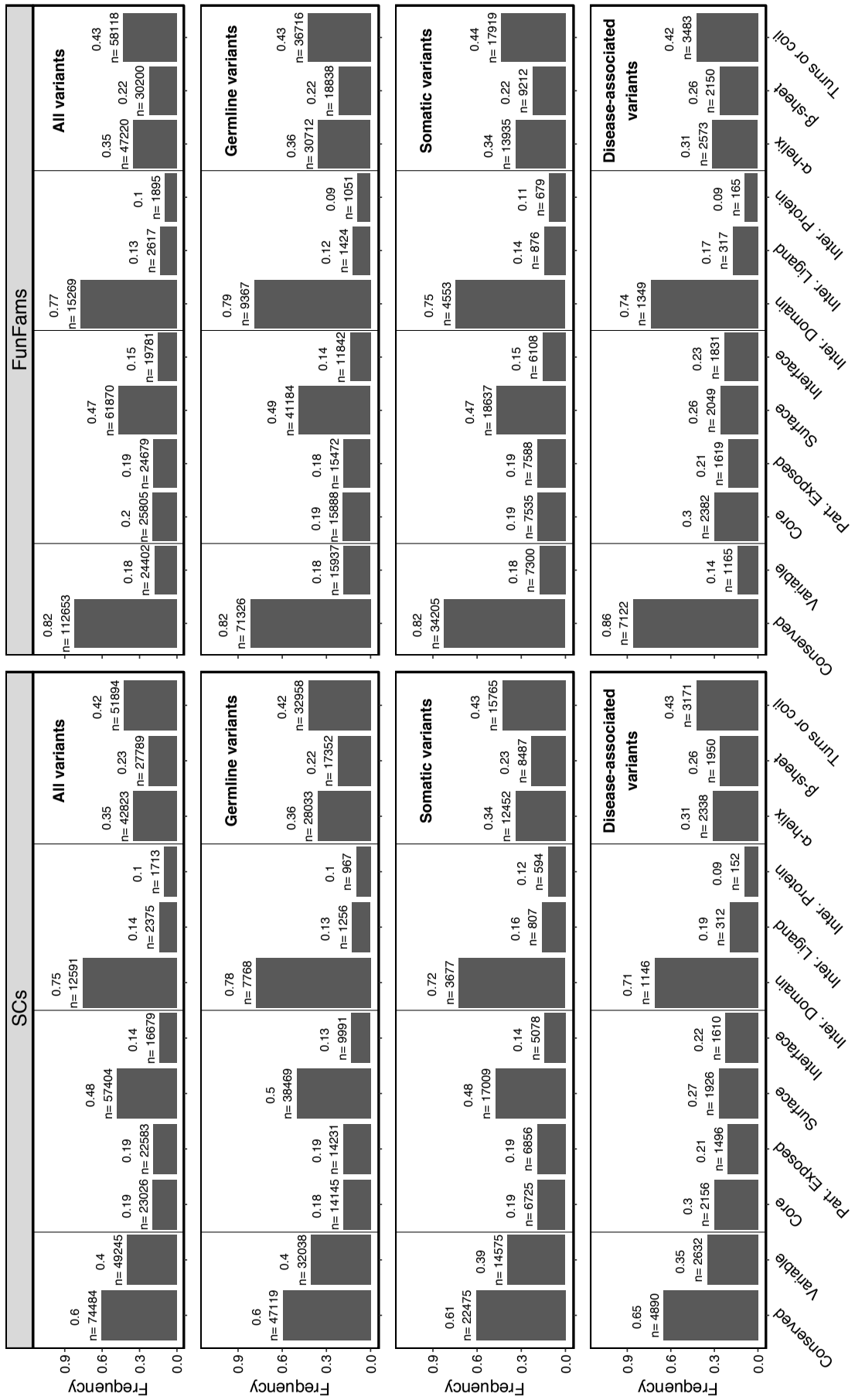

SCs
FunFams
All variants
Germline variants
Somatic variants
Disease-associated variants
Frequency
Conserved
Variable
Core
Part. Exposed
Surface
Interface
Inter. Domain
Inter. Ligand
Inter. Protein
α-helix
β-sheet
Turns or coil

Figure 4.5: Distribution of genetic variants mapped to domain members of SCs and FunFams across different structural environments. Genetic variants were grouped as germline variants, somatic mutations and disease-associated variants. The aggregation of all variants in the three classes is also provided. n corresponds to the total number of genetic variants in each class that fall into that particular structure environment. Structural environments/regions were defined as described in Section 4.3.2.

Some substantial differences can be still be seen particularly for the frequency of disease-associated variants located in the protein core, in the interface regions, as well as in the STAMP structurally conserved and variable regions. The full analysis of how different variants in the different classes map to the different regions/environments and whether there is any enrichment of a particular variant type in a particular region is explored in the next Section 4.4.3.

4.4.3 Variation odds ratio across different structural environments

In order to identify structural regions/environments that are enriched or depleted in genetic variants, a measure of the odds ratio (OR) was used to compare the proportion of variants that are mapped onto a variety of structure regions/environments (Section 4.3.2). Figure 4.6 summarises the log OR obtained when comparing variation in the three classes mapped onto different structure environments, for both CATH SC and FunFam domain families. Log OR scores are calculated for each environment and take into account both the observed number of variants and the total number of residues that compose each region. Log OR scores above 0.0 correspond to enrichment of a particular variation class in a particular environment (e.g. structurally conserved regions in the MSA), when comparing to the same class in a

different environment (e.g. variable regions).

The overall analysis of the log OR indicates that although there are differences in the log OR profiles obtained for germline and somatic variants, more dramatic differences are observed for disease-associated variants. This is true for the log OR calculated for both SCs and FunFams, but is less pronounced for the interface environments, where germline variants and somatic variants are significantly differently distributed (p-value $< 1 \times 10^{-5}$), for both SCs and FunFams. Taking the log OR profiles of SCs and FunFams together, variants in the three classes are significantly enriched (p-value < 0.01) in the STAMP structurally conserved regions (SCRs) of the domain structures when compared to variable positions. Likewise, a significant enrichment of genetic variants is observed for the Core, partially-exposed Surface, and β-sheet regions (p-value < 0.001). Germline and somatic variants are also significantly enriched at the surface, whereas disease-associated variants are depleted (p-value < 0.001).

The analysis of the log OR profiles comparing the different variation classes within each structural environment indicates that substantial differences are observed for SCs and FunFams. Considering the composition of different structure regions/environments (Figure 4.5) and the reliability assessment of the MSAs performed in Section 3.4.2, the overall results indicate that a more reliable characterisation of variants can be performed within FunFams. Additionally, there is a slightly higher number of variants in the three classes mapped to FunFams when compared to SCs (Table 4.1 and Figure 4.5). The remaining variation analysis performed in this Chapter, as well as the analyses performed in Chapter 5, were therefore performed in the context of CATH FunFam domain families.

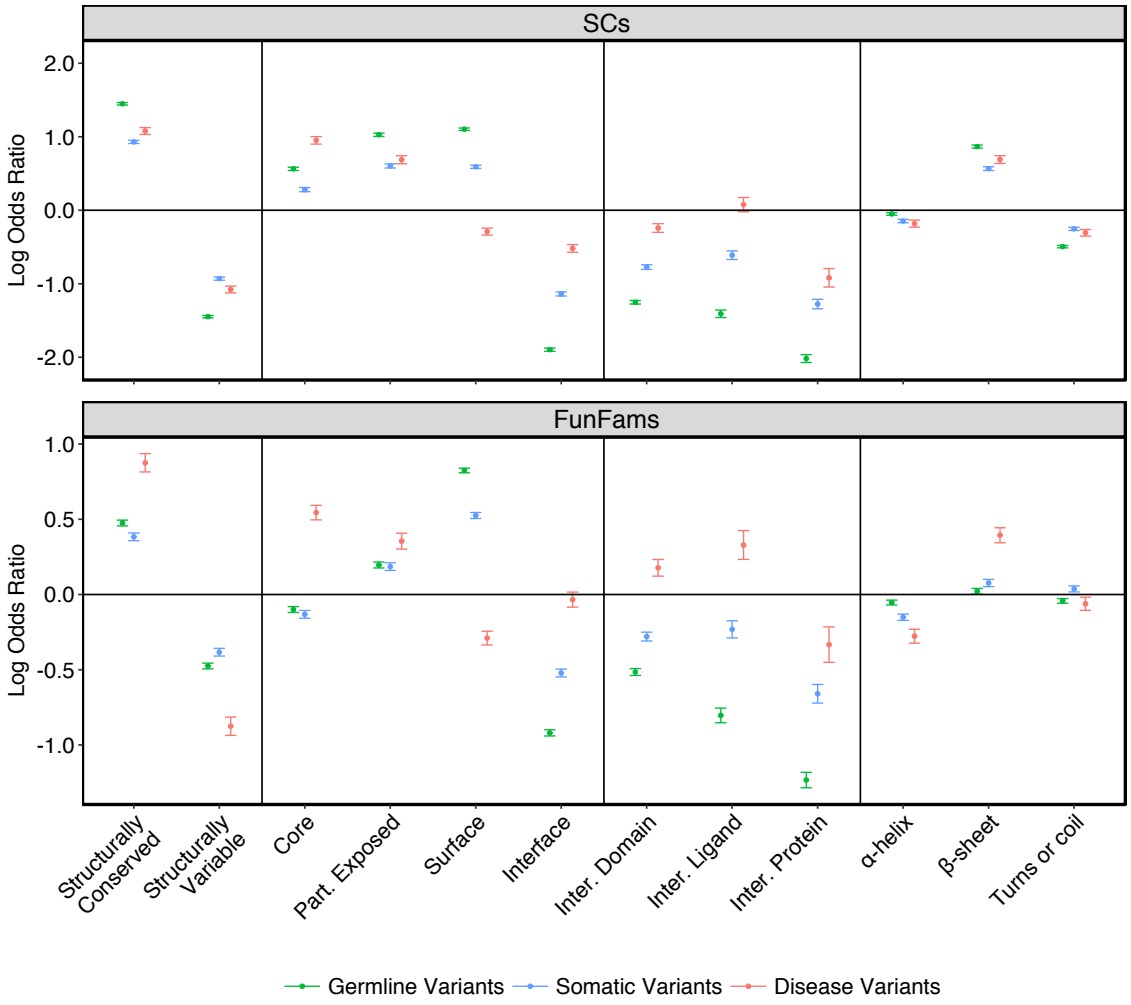


Figure 4.6: Log Odds Ratio (OR) scores obtained for the comparison of genetic variants mapped to domain members of SCs and FunFams across different structural environments. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red). The error-bars provided correspond to 95% C.I.

The comparison of the log OR between variation classes for FunFams (Figure 4.6), shows a strong enrichment of disease-associated variants at STAMP SCRs, at the protein Core, Interfaces (domain-domain and domain-ligand), and β-sheet regions (p-value $< 1 \times 10^{-5}$). This result is consistent with work reported by others, which identified a strong correlation between disease-associated nsSNPs and solvent inaccessible area and interacting interfaces (Wang and Moult, 2001; Burke et al., 2007; Sunyaev et al., 2001; de Beer et al., 2013; David and Sternberg, 2015; David

et al., 2012). The breakdown analysis of log OR scores in interaction interfaces shows that disease-associated residues are comparatively enriched in interactions with domains and ligands, when compared to domain-protein interactions. Germline and somatic variants are differently distributed at all interface environments. Somatic variants are less depleted at interfaces when compared to germline variants, which potentially accounts for the fact that some somatic variants are driver mutations in cancer (Gulati et al., 2013; Nishi et al., 2013; Alexandrov et al., 2013; Berliner et al., 2014; Cancer Genome Atlas Research Network et al., 2013; Supek et al., 2014).

4.4.4 Analysis of genetic variation amino acid exchanges

Figure 4.7 highlights how 'from' and 'to' amino acid exchanges obtained for the three variation classes compare to the frequency of each amino acid in the human proteome. Overall, some differences are observed for the frequency of amino acid mutations. Mutation 'from' Arg is significantly enriched (p-value < 0.001) for all variation classes. Minor enrichment is also observed for Asp and Gly, for somatic and disease-associated variants, respectively. Regarding mutation 'to', charged residues Asp and Glu, as well as neutral Ala and Gly, are significantly depleted (p-value $<$ 0.01), across all variation classes.

Figure 4.8 shows the frequency in which all of the 'from' and 'to' amino acid exchanges are observed in the variation dataset. The overall trend in amino acid exchange frequencies is similar when comparing different variation classes. Aromatic amino acids (Phe, Tyr and Trp) are among the least mutated, whereas charged residues (Arg, Asp and Glu) are among the most mutated residues.

For disease-associated variants, the most mutated residues are Arg, Leu and

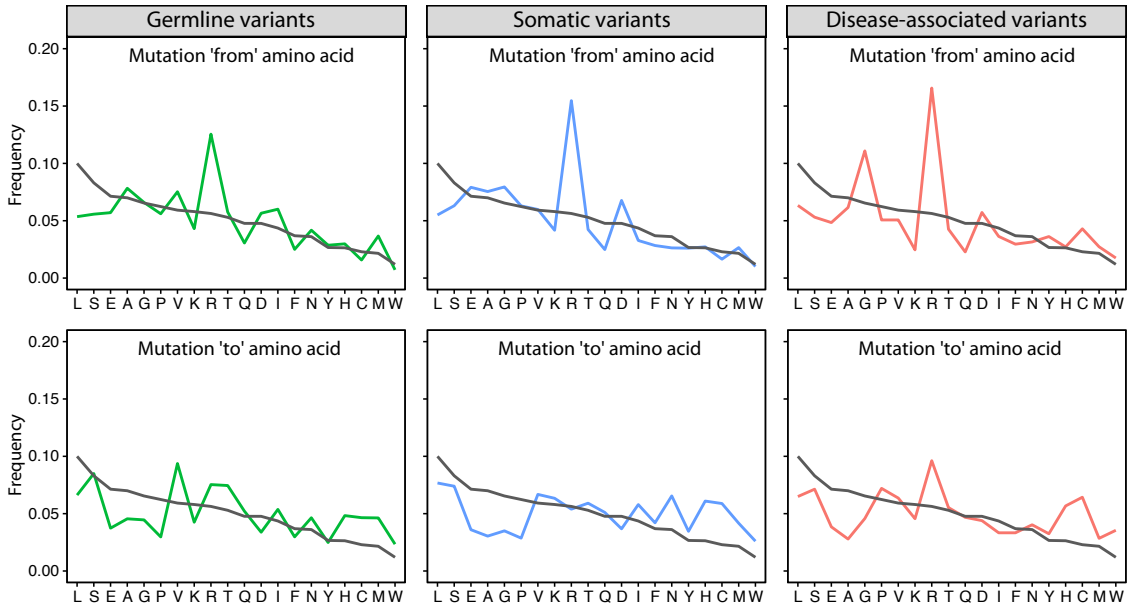


Figure 4.7: Comparison of the human proteome abundance of amino acids and mutation frequencies for all classes of genetic variants. The frequency of the observed mutations is provided for germline variants (green), somatic mutations in (blue), and disease-associated variants (red). These frequencies were compared to the frequency of each amino acid in the proteome (dark-grey line). Amino acids are arranged by 1-letter code according to decreasing abundance in the human proteome (de Beer et al., 2013).

Gly. In fact, the biggest frequency difference for 'from' mutations is observed for Gly which is found mutated more often in disease-associated variants (p-value $<$ 0.01). This likely results from the fact that Gly is the smallest amino acid and structurally dissimilar to larger aromatic and charged amino acids, adopting many angle conformations inaccessible to other amino acids (Martin et al., 2002; Bradshaw et al., 2011; David and Sternberg, 2015). The side chain of Gly is constituted by a single hydrogen atom, which can adopt a much larger range of conformations than other residues, providing structural flexibility that is lost upon mutation. Any changes from this amino acid are likely to lead to important structural changes. Cys and Arg are enriched for mutation in disease-associated variants. Ile and Val are enriched for mutation in germline variants, whereas for somatic mutations both

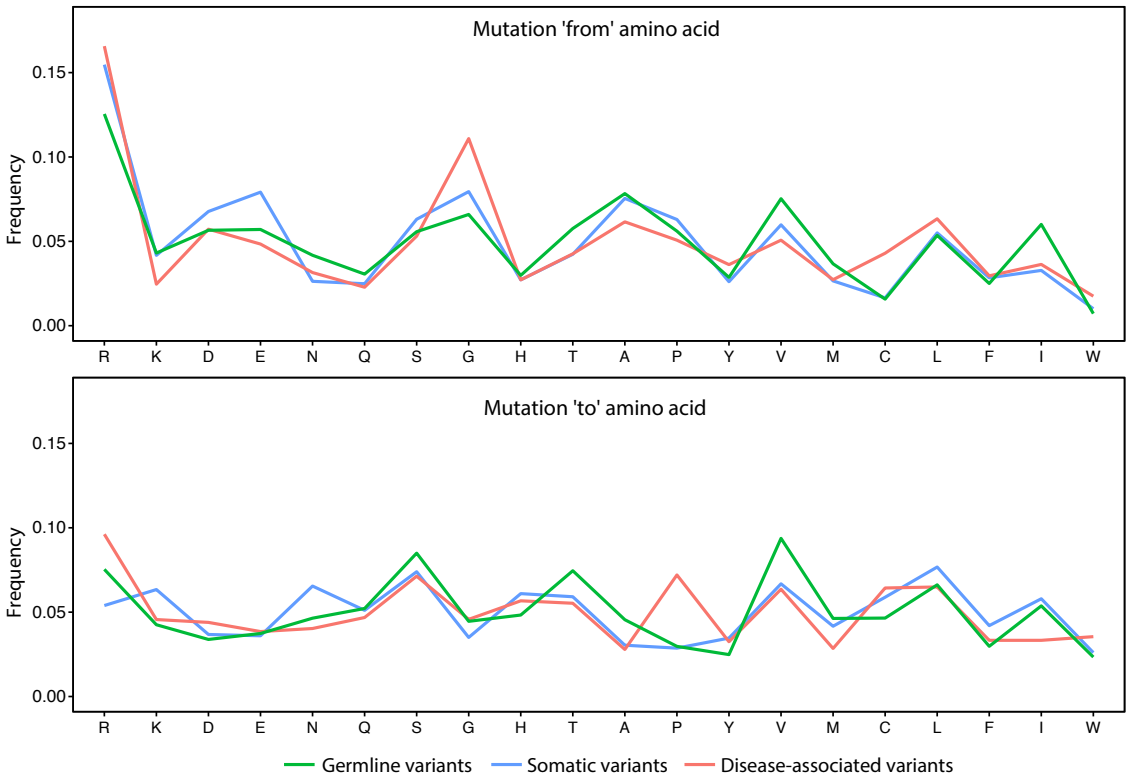


Figure 4.8: Amino acid exchanges observed for the three classes of genetic variants mapped onto FunFams. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red). Amino acids are arranged by 1-letter code according to increasing hydrophobicity (least hydrophobic is left and most hydrophobic is right) according to the Fauchère et al., 1988, scale.

acidic residues, Asp and Glu, are similarly enriched. Considering the nature of the amino acids, disease-associated exchanges enrichment for Gly and Cys (p-value $<$ 0.01) is in agreement with the fact that these amino acids are structurally dissimilar and any changes from these are likely to lead to important structural changes. In terms of amino acids whose exchanges are depleted, only Lys and Ala are marginally less mutated (p-value $>$ 0.05) for disease-associated variants. The depletion of Ala in the disease-associated variants agrees with the fact that Ala is a neutral amino acid in terms of its size and physicochemical properties.

For 'to' amino acid exchanges, some frequency differences can also be identified. Among the most dramatic differences are the enrichment of mutations to Pro and

Arg found for disease-associated variants (p-value < 0.01). Enrichment in mutations that result in Val, Thr and Ser for germline variants, as well as mutation to Asn, Lys and Leu, for somatic variants, are also observed (p-value < 0.05). The difference in the profiles obtained for the 'to' exchanges, between germline and somatic variants, indicates that the mutational profile of somatic variants is different from that of the neutral germline variants. Mutations to Met and Ile seem to be depleted for disease-associated variants which are frequent for both germline and somatic variants. Considering the nature of the amino acids, disease-associated exchanges enrichment to Arg, Pro and Cys, are in line with the expected outcome of mutation to these residues. Mutation to Arg is likely to affect the protein because of the introduction of a large charged amino acid. Similarly, mutation to Pro is known to be problematic since the introduction of Pro is likely to impose angle constraints to the protein backbone chain conformation. Pro side-chain locks its backbone dihedral angle, causing it to be exceptionally rigid conformationally. The introduction of a Pro residue is known to have the potential to disrupt secondary structure elements (Martin et al., 2002; Wang and Moult, 2001).

Figure 4.9 shows the breakdown of all amino acid exchanges observed in the different genetic variation subsets. Each subplot shows the results of mutation from a specific amino acid (e.g. Arg at top left) to every other amino acid. Interestingly, all possible variant exchanges, in terms of codon changes, were found for all amino acids. Among the more dramatic changes in the frequency of mutation are mutations from Arg, Gln, Ser, Ala and Leu to Pro, found for disease-associated variants. Smaller frequency differences are observed for the mutational profile of germline and somatic variants, when compared to those from disease-associated variants.

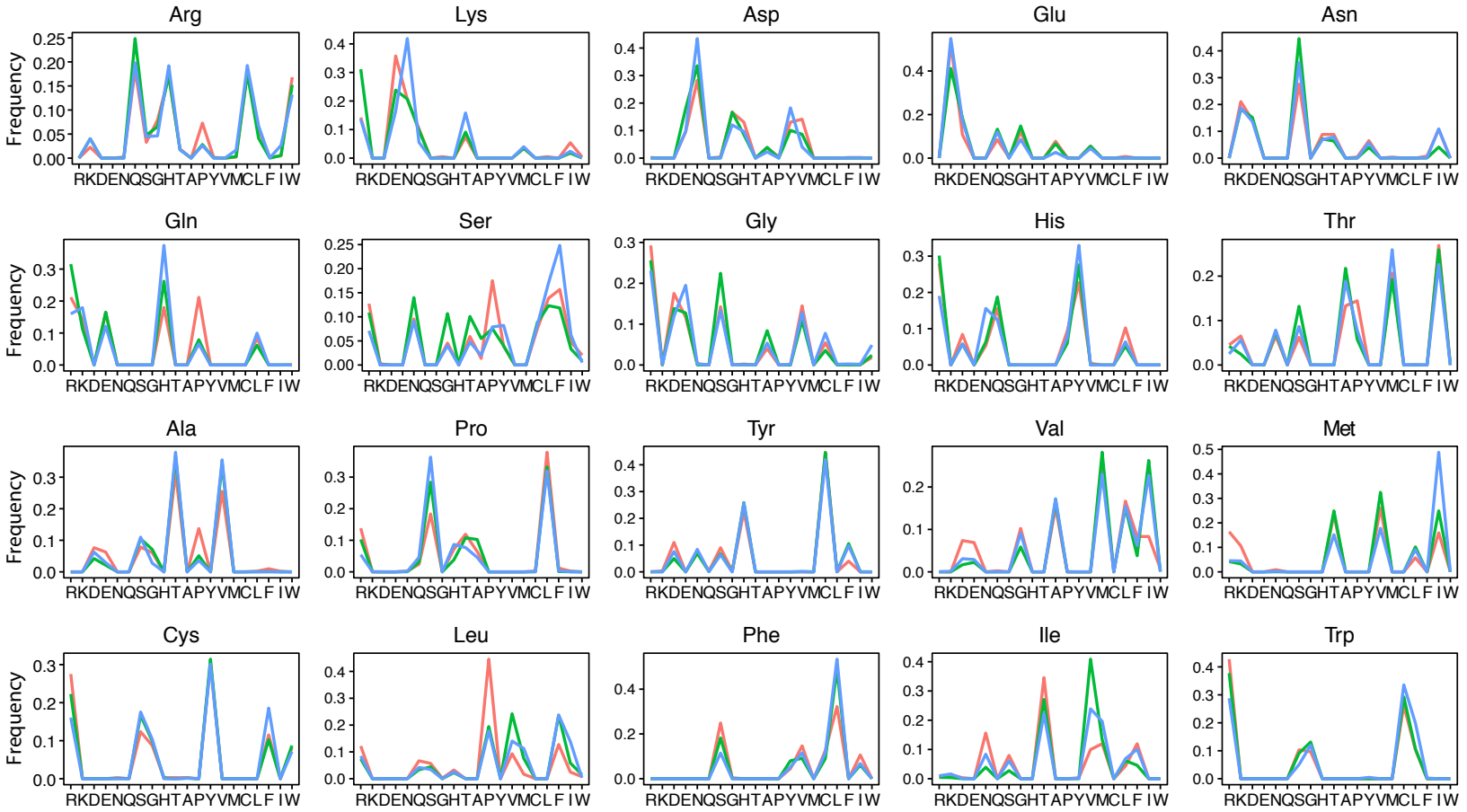


Figure 4.9: Comparison of the mutations frequencies for all classes of genetic variants. The frequency of observed mutations is provided for germline variants (green), somatic mutations in (blue), and disease-associated variants (red). Each plot shows the results of mutation from a specific amino acid (e.g. Arg at top left) to every other amino acid. Amino acids are arranged by 1-letter code according to increasing hydrophobicity (least hydrophobic is left and most hydrophobic is right) according to the Fauchère et al., 1988, scale.

Amino acid exchanges from Tyr to Cys and from Cys to Tyr are common among germline and somatic variants, while entirely depleted for disease-associated variants (Figure 4.9). These results are in agreement with those obtained for the analysis of germline variants in the 1kGP (The 1000 Genomes Project Consortium, 2012), and disease-associated variants from OMIM (Amberger et al., 2015), which are both a part of the variant datasets studied here, as performed by de Beer et al., 2013.

In order to obtain a summary view of the entire mutational space and to calculate the rate of exchange for genetic variants in the three variant classes, log-odds mutability matrices were generated by calculating the log-odds scores of the observed mutation frequencies divided by the expected mutation frequencies (Equation 4.2). Figure 4.10 shows the log-odds mutability matrices for all amino exchanges in the three variation classes, as well as difference matrices which highlight the differences between the variation datasets. Unlike other substitution matrices such as BLOSUM62 (Henikoff and Henikoff, 1992), which assume that the mutation direction is unknown and produce symmetric matrices, here the depth of the variation data allows matrices to be generated with embedded directionality. The relevance of these matrices is three-fold: 1) they capture the rate and exchangeability of missense variants, cancer mutations and disease-associated mutations separately; 2) which can be used to model the probability of fixation of different mutation classes in protein evolution; 3) as well as to aid the prediction of the effects of amino acid exchanges. The analysis of the extreme differences observed for the mutation matrices of germline, somatic and disease-associated variants, highlights the differences observed and discussed previously (Figure 4.10 D-F).

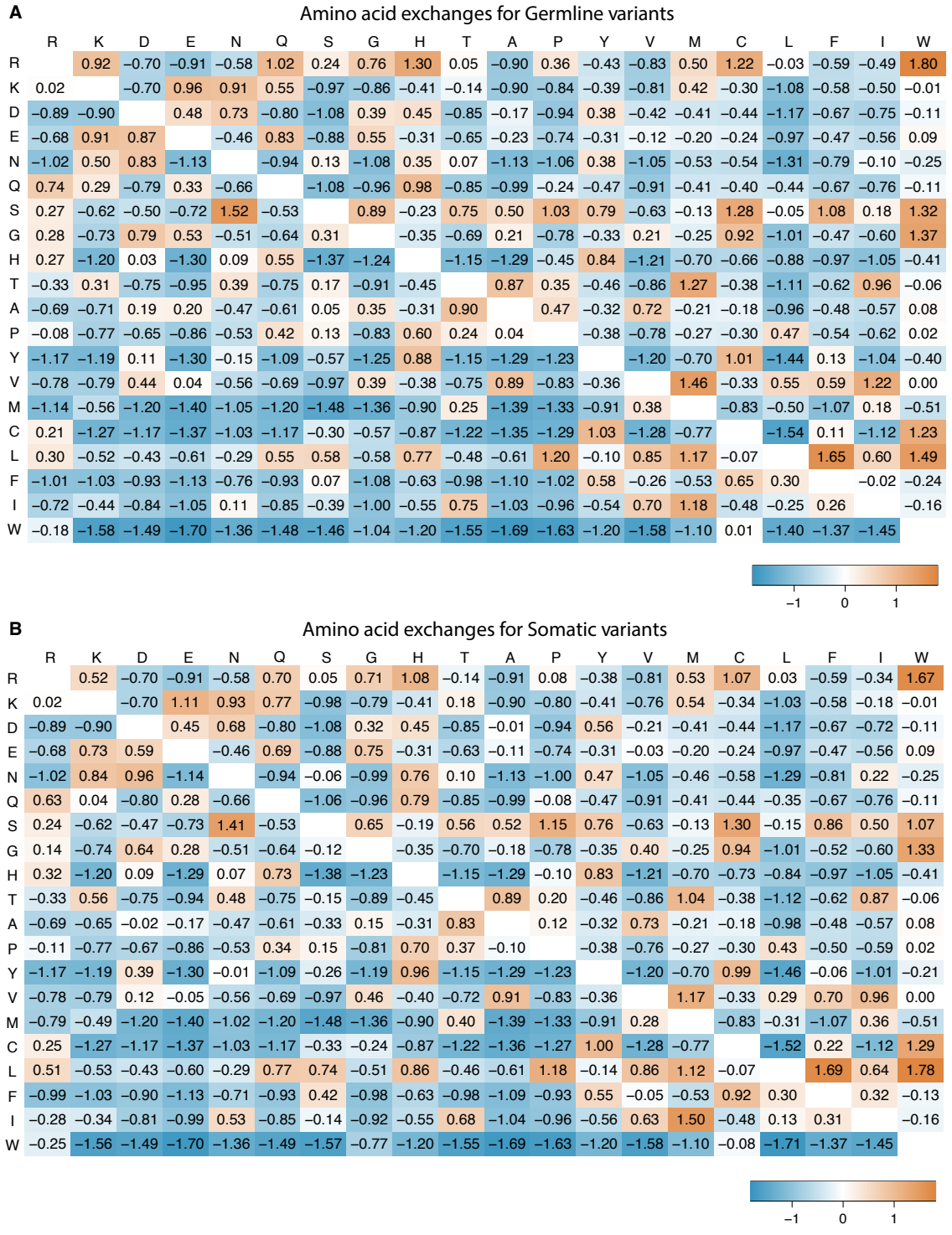


Figure 4.10: Log-odds mutability matrices for amino acid exchanges observed for the three genetic variation classes. Exchange matrices are provided for: A) germline variants; B) somatic variants; C) and disease-associated variants. The comparison (difference) matrices are provided for: D) germline and somatic variants; E) germline and disease-associated variants; and F) somatic and disease-associated variants. Amino acids are arranged by 1-letter code according to increasing hydrophobicity (least hydrophobic is left/top and most hydrophobic is right/bottom) according to the Fauchère et al., 1988, scale.

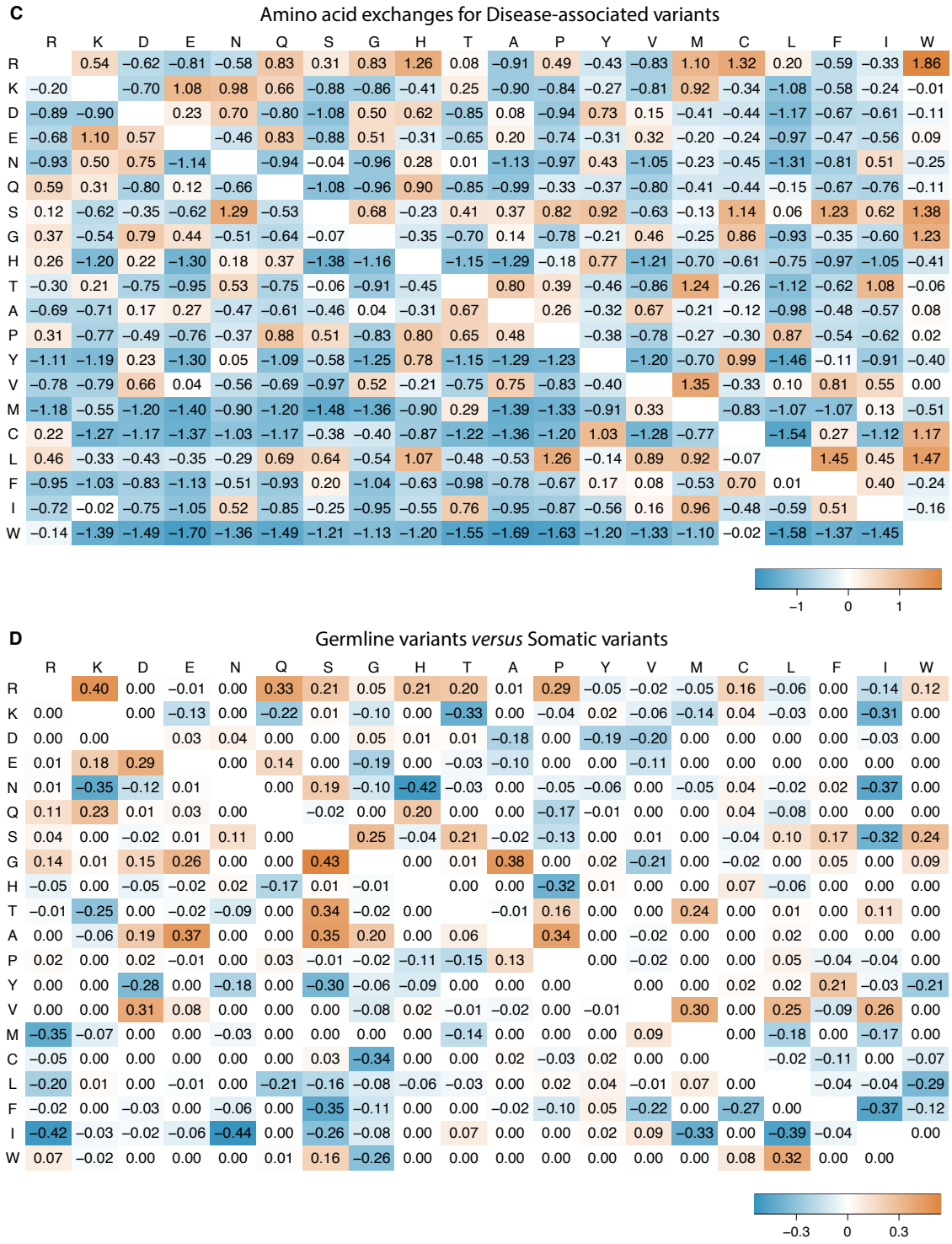


Figure 4.10
(cont.)

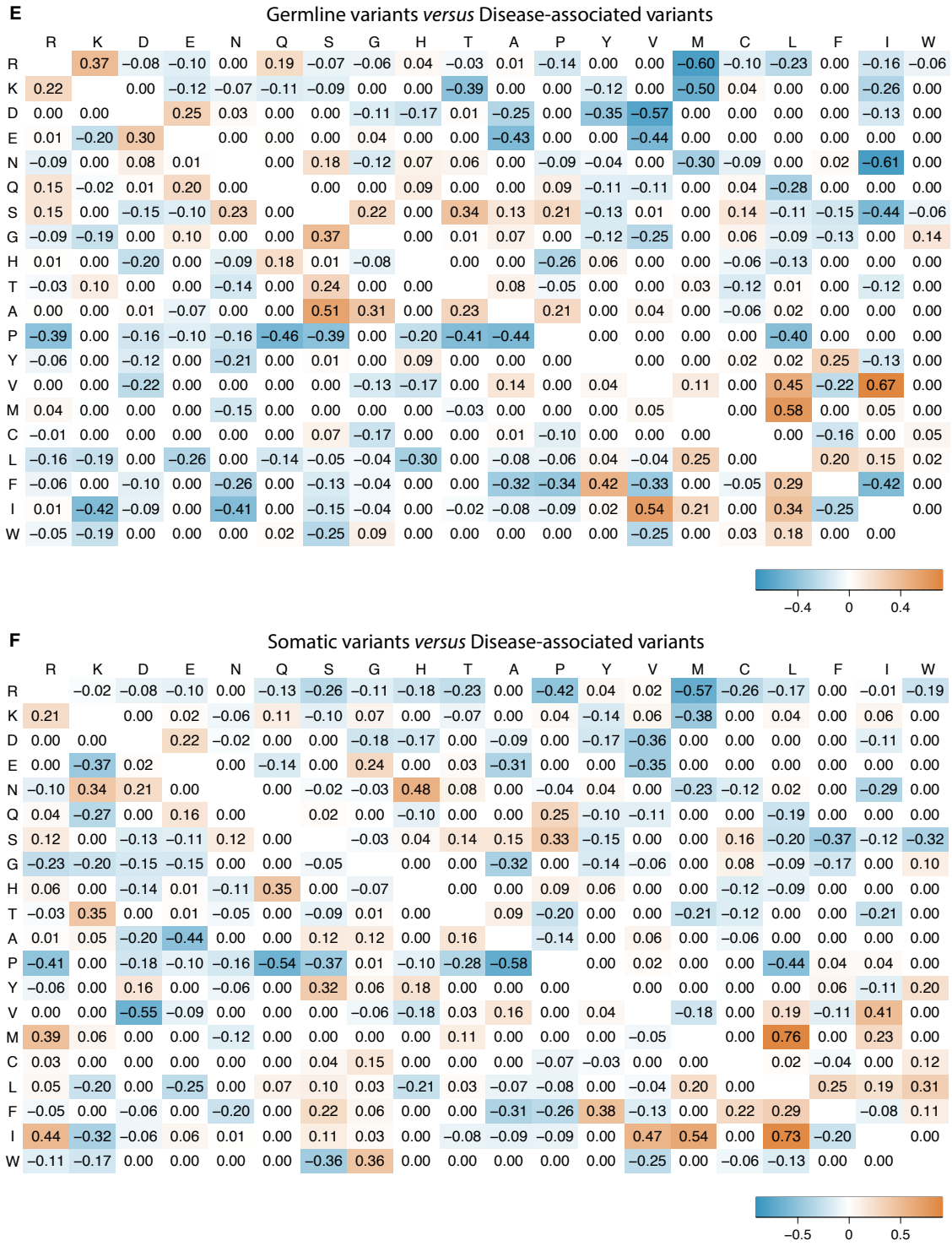


Figure 4.10
(cont.)

4.4.5 Analysis of genetic variation exchanges according to physicochemical properties

In order to further investigate the amino acid exchanges observed within the variation dataset, amino acids were grouped according to their physicochemical properties. Two single state alphabets, 'Chemical_A' and 'Chemical_B', which organise each amino acid according to their unique physicochemical properties, were investigated. Figure 4.11 shows the frequency in which all of the 'from' and 'to' physicochemical group exchanges are observed in the variation dataset. The analysis of exchanges for both physicochemical alphabets reveals that only minor frequency differences (not significant) are observed for variants in different variation classes. In agreement with the amino acid exchange profiles (Figure 4.9), Gly and Pro seem to be enriched for mutation in disease-associated variants. There is also a small enrichment of mutations occurring at hydrophobic and acidic residues (p-value < 0.05), for germline and somatic variants, respectively. A minor enrichment of mutation to hydroxy-group containing residues (Ser and Thr) is also observed for germline variants (p-value > 0.05).

Figure 4.12 shows the breakdown distribution of physicochemical changes introduced by genetic variants according to the two alphabets defined. Regarding Chemical_A, the most dramatic transition is observed for exchanges between hydrophobic residues and Pro for disease-associated variants. Smaller changes in the frequency of exchanges between acidic residues and both hydrophobic and basic residues are also observed for somatic variants (Figure 4.12 A). Since these residues are often involved in interaction, this might affect the propensity to interact and

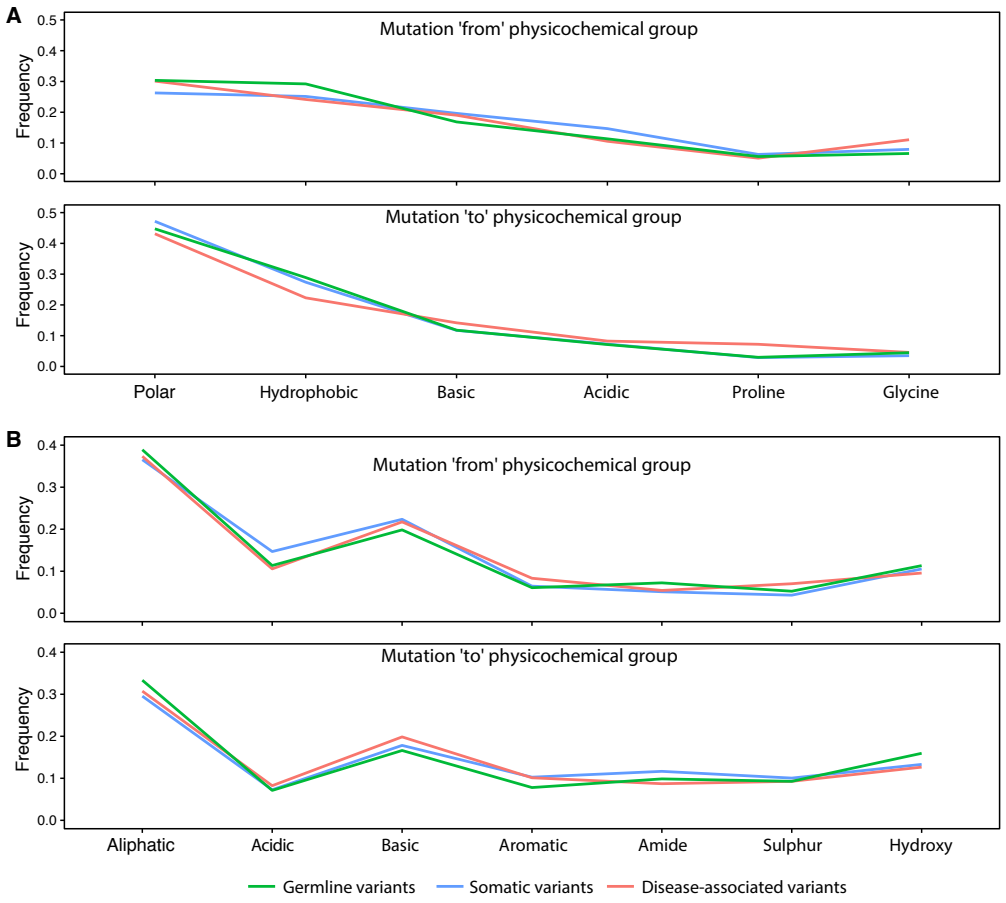


Figure 4.11: Physicochemical exchanges observed for the three classes of genetic variants. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red). Amino acids were grouped into two alphabets: A) 'Chemical_A' and B) 'Chemical_B'.

lead to changes in binding partners with regulatory implications. The analysis of physicochemical transitions shown for the Chemical_B alphabet indicates that mutation of aliphatic residues such as Leu, Ala, and Val, to either acidic and basic residues, are enriched for disease-associated variants. This is expected since the introduction of charged groups is likely to affect intermolecular interactions between amino acids. Interestingly, mutation of residues containing sulphur (Met and Cys) also show some enrichment (p-value < 0.05) in exchanges to basic and aromatic residues, for disease-associated variants (Figure 4.12 B). Cys often participates in disulphide bonds which are disrupted when these residues are mutated.

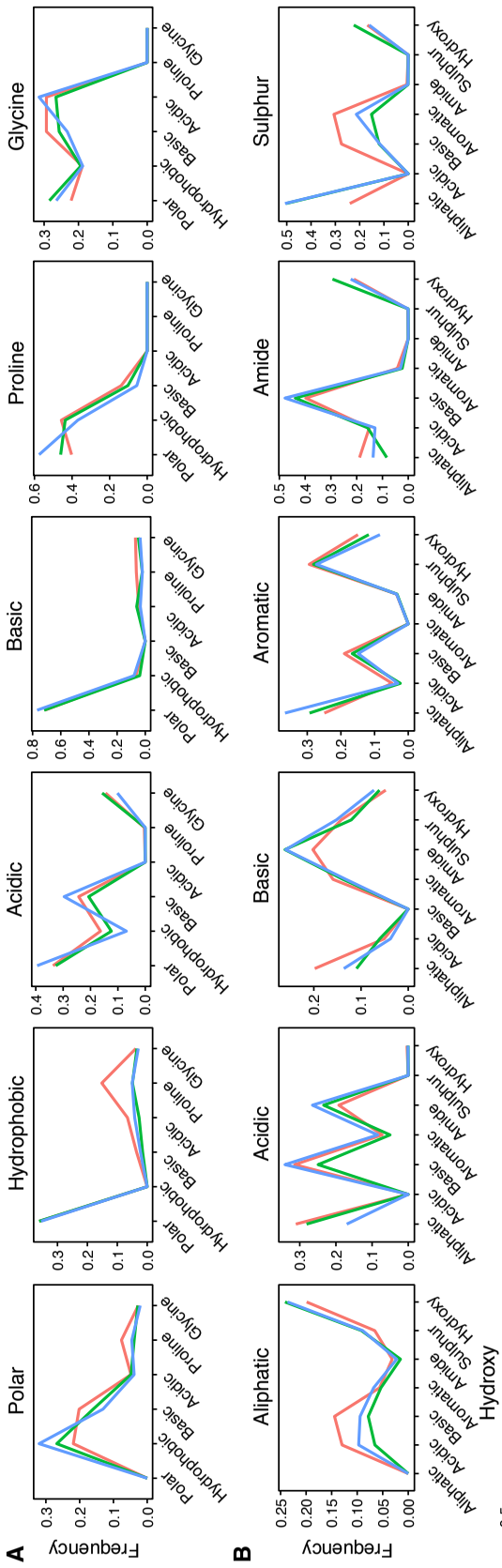


Figure 4.12: Comparison of the physicochemical exchange frequencies for all classes of genetic variants. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red). Amino acids were grouped into two alphabets: A) 'Chemical_A' and B) 'Chemical_B'.

The biggest frequency differences observed between somatic variants and germline variants are observed for exchanges from Met and Cys residues (sulphur group) to chemically complex aliphatic residues, as well as for exchanges from hydroxy-group containing residues (Ser and Thr) to aromatic residues (Figure 4.12). Mutation from Ser and Thr is likely to have consequences, especially since these residues are often post-translationally modified (e.g. phosphorylation and O-linked glycosylation).

Figure 4.13 explores the overall variant transitions in terms of amino acid properties including: residue's atomic mass (Knapp, 1996); average volume (Zamyatnin, 1972); and hydrophobicity (Fauchère et al., 1988). The differences between the standard amino acid values for mass, volume and hydrophobicity for the 'from' and 'to' amino acids were calculated and plotted as the density of differences. The comparison of the hydrophobicity exchange profiles obtained for the three variant classes reveals that, for disease-associated variants, transitions are spread slightly more widely, to the extremes of the scale. This means that disease-associated transitions lead to bigger changes in hydrophobicity. Only minor differences are observed between the profiles of germline and somatic variants. Nevertheless, germline variant transitions are likely result more frequently in hydrophobic-equivalent amino acids.

A similar trend is observed for the amino acid transition profiles which are related to amino acid size (atomic mass and average volume). Germline variants and somatic variants are often mutated to identically sized amino acids when compared to disease-associated variants, which lead to more extreme amino acid size exchanges. In agreement with previous observations (de Beer et al., 2013), these results indicate that drastic amino acid transitions are likely to disrupt stability/activity of proteins and lead to disease states (Miller and Kumar, 2001; Tang et al., 2004; Stone and

Sidow, 2005; Khan and Vihinen, 2007; Briscoe et al., 2004).

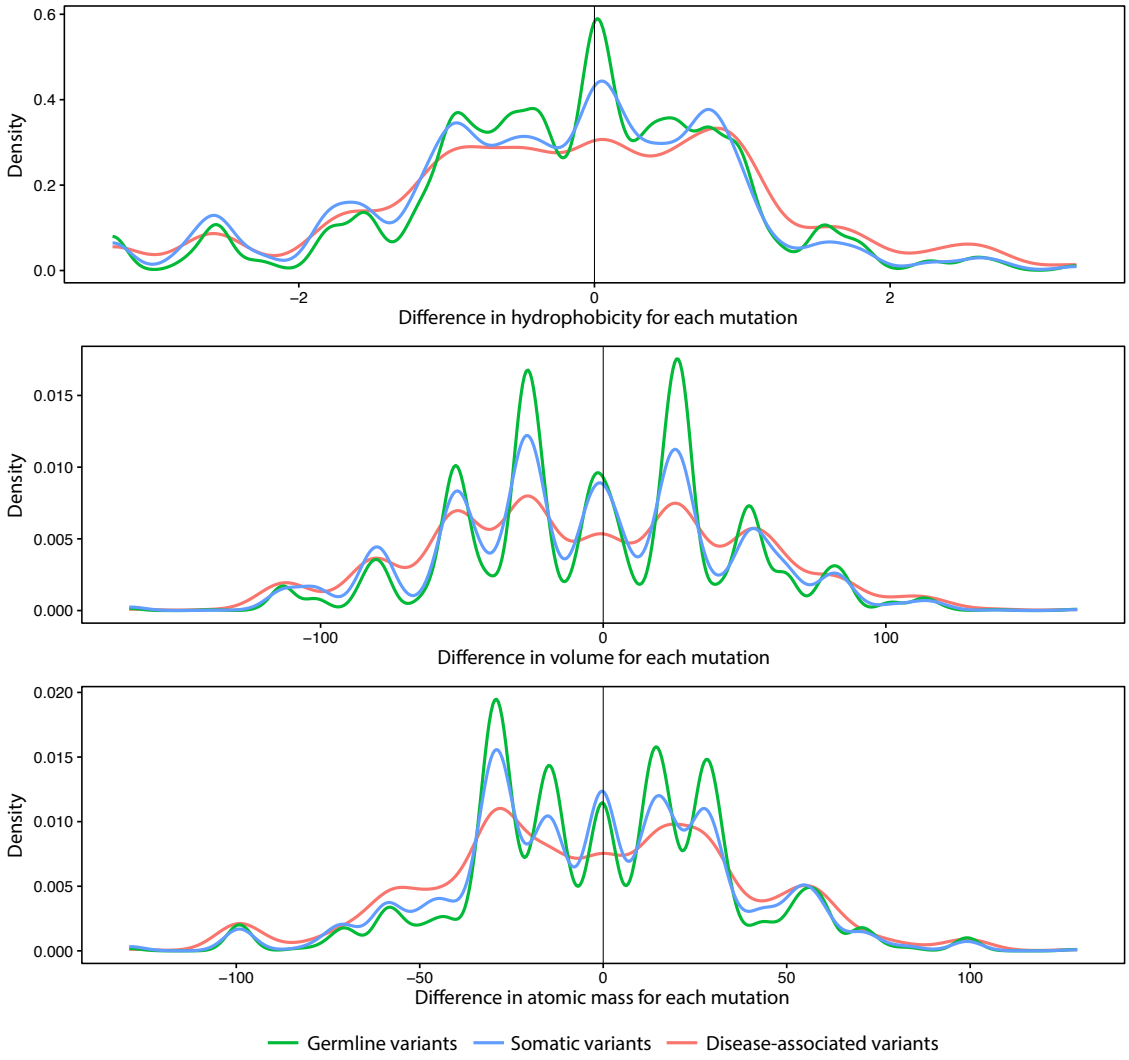


Figure 4.13: Genetic variation transitions in terms of amino acid hydrophobicity, volume, and atomic mass. Amino acid transitions were calculated as the difference between the property (i.e. hydrophobicity, volume or mass) value of the original (mutated) amino acid and that of the resulting amino acid. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red)

4.4.6 Analysis of conservation

Various methods to measure conservation in MSAs have been developed (Valdar, 2002; Johansson and Toh, 2010). Some of these methods take evolution models and

residue background distribution into account, but often result in complicated interpretation of the conservation scores. Here a simple measure of conservation was used to compare the degree of conservation among MSA positions (columns) to which genetic variation could be mapped (Figure 4.1). Shannon's Entropy (H) is one of the simplest conservation measures, that only takes into account the amino acid frequencies in the MSA columns (Shannon, 1948). Figure 4.14 shows the distribution of Shannon's entropy conservation scores across the three different classes of variation. Previous work has shown that disease-associated variants are often located within conserved regions of proteins (Hu et al., 2000; Guharoy and Chakrabarti, 2010; Fiser et al., 1996). Interfaces have been shown to be more conserved than the rest of the surface (Nooren and Thornton, 2003b; Bordner and Abagyan, 2005). In agreement, there is a substantial proportion of disease-associated variants that are found in highly conserved regions (low H score). In fact, both germline and somatic variants follow the same trend of being located in conserved positions of the alignment for FunFam domain families.

All genetic transitions in the variation dataset were investigated within the full

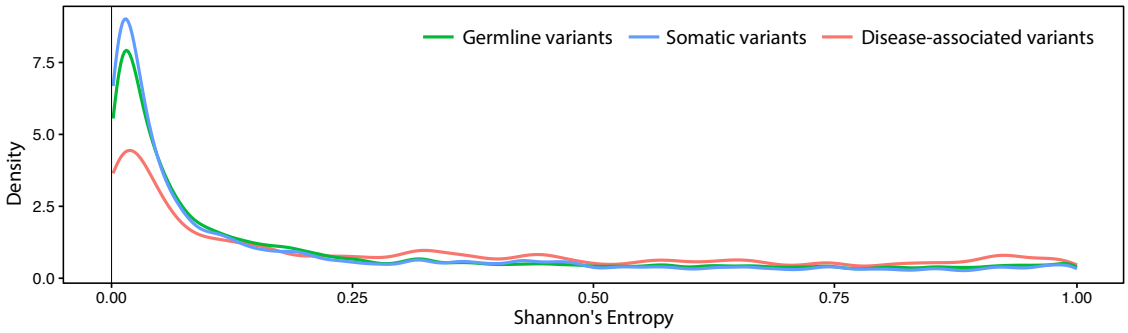


Figure 4.14: Distribution of Shannon's entropy conservation scores obtained from the alignment columns to which genetic variants could be mapped. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red).

MSAs to assess whether the amino acids resulting from mutation 'to' are already found in the aligned positions. Table 4.5 shows the number of variant exchanges present in the alignment column where they map to. Interestingly, although 31% of the resulting germline and somatic variant amino acids are observed in the alignment column to which the variant in analysis is mapped to, this number drops to 24% for disease-associated variants. This decrease is expected since disease-variants are preferentially found in conserved positions (Figure 4.6). This result is in agreement with the observation that mutation of conserved positions is likely to affect protein function. Interestingly, the disease-associated variants which are observed might result from misalignment of the structures/sequences, but most likely from particular features of the domain for which the disease variant is observed and that confer functional uniqueness. These include unique domain or ligand interaction partners as well as unique structural/functional features that might be disrupted in this particular family domain member. Additionally, an event such as the co-evolution of interface residues in homologous proteins is also a possibility (Halperin et al., 2006; Marks et al., 2012; Pazos and Valencia, 2008).

Table 4.5: Summary of the number of variant exchanges for which the resulting amino acid is already present in the aligned position in the MSA.

Class of variants	**Observed exchange**	**Total variants**
All variants	42,243 (0.31)	137,060
Germline variants	27,273 (0.31)	87,268
Somatic variants	12,976 (0.31)	41,505
Disease variants	1,994 (0.24)	8,287

4.4.7 Predicting the consequences of genetic variation

Prediction scores from Polyphen-2 (Adzhubei et al., 2010) and SIFT (Kumar et al., 2009) were collected for each variant. Polyphen-2 predicts the effect of an amino acid substitution on the structure and function of a protein using sequence homology, Pfam (Finn et al., 2014a) annotations, protein structures, secondary structure states, and several other databases and tools. The Polyphen score represents the probability that a substitution is damaging, so prediction scores (P) values nearer one are more confidently predicted to be deleterious. SIFT predicts whether an amino acid substitution is likely to affect protein function based on sequence homology and the physicochemical similarity between the alternating amino acids. The SIFT score is the normalised probability that the amino acid change is tolerated so that scores nearer 0 are more likely to be deleterious (note that this is the opposite to Polyphen-2). Categorical classification states are derived from P resulting in three states for Polyphen-2: benign, possibly-damaging, and probably-damaging; as well as two states for SIFT: benign and tolerated. SIFT and Polyphen2 were chosen because prediction scores were provided by the Ensembl and UniProt Variant APIs.

Figure 4.15 shows the distribution of Polyphen-2 and SIFT categorical states observed across the three variant classes. The distribution of Polyphen-2 scores indicates that there is a high number of germline and somatic variants classified as deleterious. Interestingly, according to the Polyphen-2 categorical states, a higher proportion (68%) of disease-associated variants is classified as probably-damaging, when compared to germline and somatic variants. This result is expected since the methods have been trained to classify variants by considering physicochemical

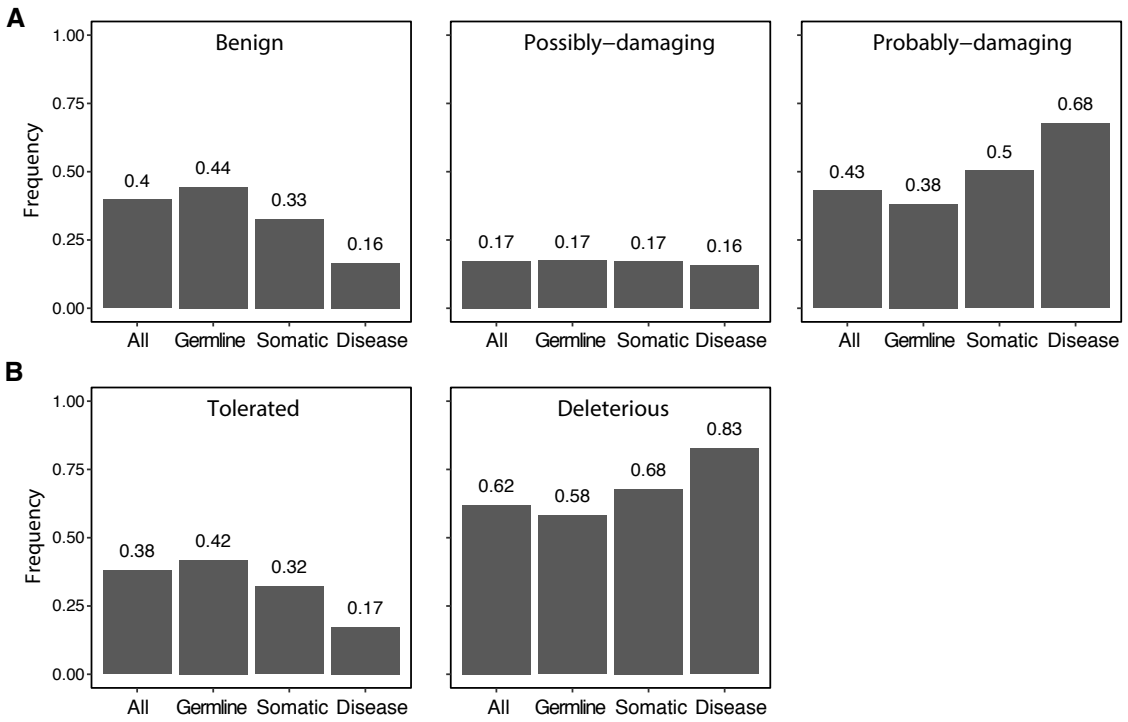


Figure 4.15: Bar-plot showing the categorical SIFT and Polyphen-2 predictions observed across the three variation classes. Bar-plot showing the: A) Polyphen-2 (Adzhubei et al., 2010), and B) SIFT (Kumar et al., 2009) categorical transitions observed across the classes of variants. Qualitative categorical prediction states are defined from the prediction scores (P) provided by SIFT as: deleterious ($P \leq 0.05$); or tolerated ($P > 0.05$). Similarly, categorical states generated for Polyphen-2 (HumVar) are classified as: benign ($P < 0.446$); probably damaging ($0.446 \leq P < 0.909$); or possibly damaging ($P \geq 0.909$).

transitions and conservation. Besides, some of the disease-associated variants might be included in the training sets used to train Polyphen-2. Germline and somatic variants are enriched within benign classification states from Polyphen-2 (44% and 33%, respectively). Figure 4.16 shows the distribution of Polyphen-2 and SIFT prediction scores across the different genetic variation classes. In agreement with the distribution of Polyphen-2 scores observed, an overall lower proportion of variants across the three classes is classified as possibly-damaging (16-17%) (Figure 4.15). The profiles obtained for Polyphen-2 indicate that the method is biased towards the extreme scores, where a variant is either classified as benign or probably-damaging.

SIFT categorical classification seems to be in agreement with Polyphen-2 predictions, classifying 83% of disease-associated variants as deleterious. Since SIFT also predicts a high proportion of germline and somatic variants to be deleterious (38% and 50%, respectively), it is likely that the number of false positive predictions is high and biased towards predicting deleterious (low P) mutations (Figure 4.16).

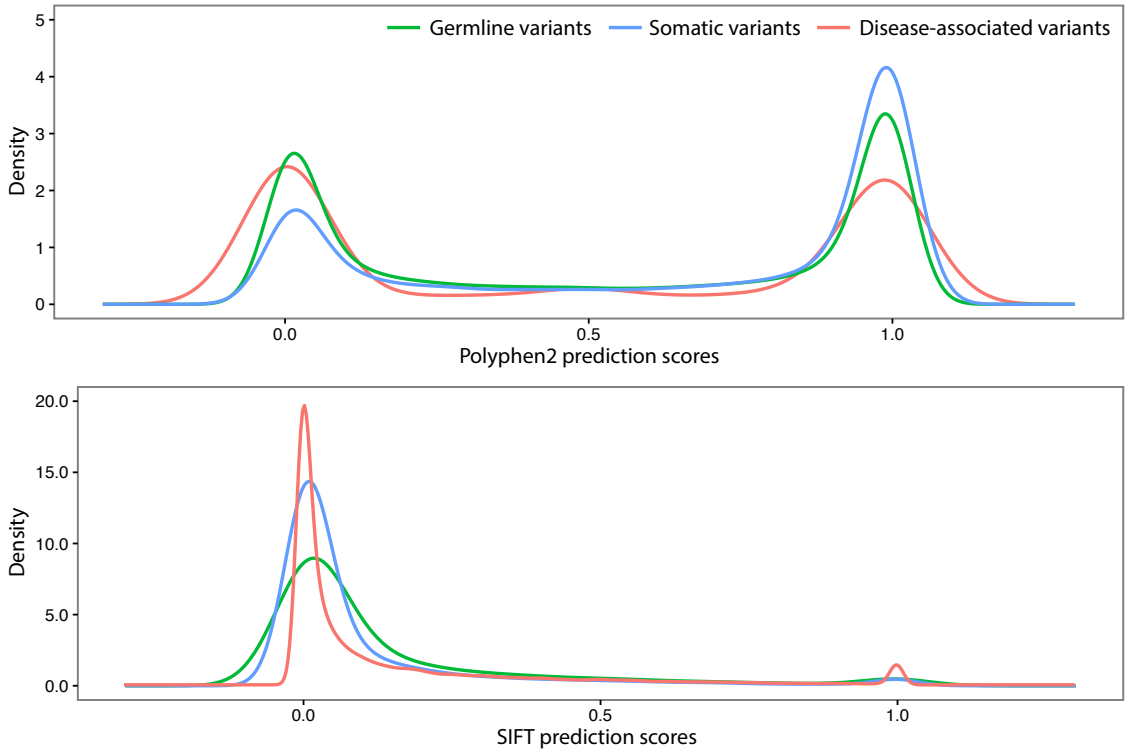


Figure 4.16: Distribution of Polyphen-2 and SIFT prediction scores for genetic variants stratified by variation class. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red).

4.5 Conclusions

This Chapter focused on the overall analysis of genetic variants in protein families and in the context of various structural environments. The approach of characterising in depth the variants across different environments by type of substitutions and annotation is important as it enables finding transitions that might be implicated

in disease, and that can potentially affect protein stability, activity and function. The global trends identified are also important for selecting priority variants and protein families for further analysis. Analysis of the genetic variation that maps onto domain-domain and domain-ligand interaction interfaces is further explored in Chapter 5.

The main conclusions of the work presented here are:

- Genetic variants were obtained from multiple sources and organised according to their annotations.
- A large number of genetic variants were mapped to protein domains in SCs and FunFams.
- 64% of the variants analysed are germline variants, which are thought to be neutral, but may contain variants unknown to cause disease.
- 30% of the variants were identified in cancer samples and were annotated as somatic variants.
- 6% of the variants are organised in OMIM, Humsavar and ClinVar, and have been annotated as disease-associated.
- Only between 3-6% of frameshift, stop-lost and stop-gained variants are annotated as disease-associated.
- Domain-domain interaction interfaces account for 63-67% of all interactions.
- 15% of interface residues participate in domain-ligand interactions.

- There is a positive correlation between the number of residues that compose different structural environments/regions and the number of genetic variants mapped to those regions.
- The bigger differences in the odds ratio are observed for disease-associated variants that map onto the protein core, surface and interface regions, when compared to neutral and somatic variants.
- Disease-associated variants are enriched on domain-domain and domain-ligand interaction interfaces, as well as in STAMP structurally conserved regions, for FunFam domain families.
- A slight enrichment of disease-associated variants is also observed within β-sheets.
- Mutability matrices were generated for all amino acid exchanges in the variant classes.
- Amino acid exchanges were analysed according to physicochemical properties.
- The more drastic physicochemical transitions are observed for mutation from and to Gly and Pro residues.
- The extremer size and hydrophobicity amino acid transitions are observed for disease-associated variants.

Chapter 5

Exploring variation at domain interfaces

5.1 Summary

This Chapter focuses on the analysis of genetic variants at interfaces. To identify significant genetic variation trend differences, variants at the interfaces were compared to variants mapped to other structural environments/regions. Variant analysis was performed at structurally conserved interface positions in the context of CATH FunFams. Analysis of domain-domain interactions by interface orientation type was also performed. The overall variation analysis follows the same general trends observed when investigating genetic variants across all domain sites. The comparison of the variation profiles in terms of amino acid exchanges and physicochemical properties obtained for different classes of variants was performed. The most important features were identified to help prioritise the analysis of the potential consequences of

variants. In particular, trends observed for disease-associated variants at the interfaces and across all structure environments were used to help investigate unclassified variants thought to be neutral. A set of protein families and domains were selected and further investigated regarding their potential effects and involvement in disease. The mechanisms in which neutral germline and somatic variants mapped to these domain interfaces might lead to stability/activity functional disruption were investigated.

5.2 Introduction

Proteins have evolved in a way that fine-tunes interactions, forming specific molecular complexes and participating in concerted interaction networks (Russell et al., 2004; Codoñer and Fares, 2008; Pazos and Valencia, 2008). Disruption of protein interactions as the result of genetic variation at the single amino acid level prompts immediate functional consequences, which may result in disease. Understanding how genetic differences affect protein-protein and protein-ligand interactions is essential to understand their effects in susceptibility to particular diseases and drug response (Giacomini et al., 2007; Lahti et al., 2012; Ma and Lu, 2011).

Park et al., 2001, found that there is a difference between domain-domain interactions that occur between two domains on the same polypeptide chain (intramolecular interactions) and those seen between domains on different chains (intermolecular), the main difference being functional. Proteins encoded on separate genes tend to be involved in modulating the flow through a metabolic pathway as they can be independently regulated. However, domains that are part of the same

enzyme, oligomer or protein tended to be involved in intra-molecular interactions as co-regulation and co-expression are necessary. The small percentage of domains (∼20% of the overall domain interactions observed) which were seen to interact in both an intra-molecular and inter-molecular manner were mainly domains of enzymes involved in small-molecule or macromolecular metabolism.

The geometry of domain combinations, specifically the order in which domain combinations occur in nature was investigated by Bashton and Chothia, 2002. They found that if domains from Superfamily A and Superfamily B were observed fused together on a single protein chain, then they could either occur in either order, e.g. AB or BA. However, only 2% of the time did both variations occur leading them to conclude that domain order is likely to be conserved because recombination of the domains has only occurred once during evolution.

Littler and Hubbard, 2005, investigated the promiscuity of intra-chain domain-domain interactions, focusing on the surface area used to form these interactions, and the relative orientations of their domain partners. The authors found that different Superfamilies show a range of promiscuity using either the same interacting surface region or by utilising several areas of their accessible surface and that interactions between two Superfamilies tended to be in the same orientation. They also found that domain interfaces are more conserved in sequence than the rest of the accessible surface of the protein domain.

5.2.1 Genetic variation at interaction interfaces

Protein-protein interactions have been combined with SNP datasets in various studies to predict the effects of mutations on protein-protein interactions and to identify

novel disease genes and correlations between diseases. Many studies on variation at interfaces have focused on the features of interaction networks rather than the structural effects of variation. Jonsson and Bates, 2006, investigated the network features of genes using a human cancer genes dataset from COSMIC (Forbes et al., 2015) and interactome data. Cancer proteins tended to have twice as many interaction partners as non-cancer proteins. In addition, the authors also found that cancer-related proteins tend to reside in large clusters, unlike non-cancerous proteins. Later, Nishi et al., 2013, showed that cancer-associated missense mutations alter binding properties of proteins ultimately affecting their interaction networks (Nishi et al., 2013). Accordingly, genes with mutations directly affecting protein-binding properties are preferentially located in central network positions, which may influence critical nodes and edges in signal transduction networks.

Goh and Choi, 2012, created a network of disease-to-gene associations, where each genetic disease is connected to the genes known to cause it, using the OMIM database. The disease-associated genes were shown to have high tendencies to interact with one another, and they were often co-expressed. In addition, the authors found that the disease-associated genes often have a similar classification in all three domains (the biological process, the molecular function, and the cellular components) regarding their GO term annotation (Harris et al., 2004). However, the study found evidence for links between high numbers of protein-protein interactions and disease-associated genes. Schuster-Böckler and Bateman, 2008, used a combination of SNPs, interactome and structural data to create a method for improving the identification of disease genes. Using the structural interactome data from iPfam (Finn et al., 2014b), Schuster-Böckler and Bateman identified residues

that made contacts between distinct polypeptide chains. The authors developed an algorithm based on conservation score and the OMIM dataset to identify which residues had disease-causing effects. From 264 proteins, 1,428 SNPs were predicted to affect protein-protein interactions.

Teng et al., 2009, used dbSNP, BLAST (Altschul et al., 1990) and the MMDB (Madej et al., 2014) to construct 3D models of protein-protein complexes with known nsSNPs in their interfaces. They argued that any mutation that occurs at the protein interface could affect the binding affinity by perturbing a normal interaction. The CHARMM program was used to examine if the binding energy of the PPIs was affected by the substitution. The nsSNPs were grouped into categories based on how they affected the energetics of PPIs. According to Teng et al., 2009, the physicochemical properties of nsSNPs alone were not enough to predict their effect on PPIs, because changes in physicochemical properties had minimal effects on binding energy. Similarly, substitutions at non-conserved regions resulted in minimal effects on binding affinity compared to those in highly conserved regions.

A recent report suggested that protein-protein interaction sites are hot-spots for disease-associated variants (David et al., 2012). By sub-grouping variants into neutral and pathogenic classes and performing enrichment analysis, the authors showed that disease-causing nsSNPs that do not occur in the protein core are more likely to be located at the interface region rather than on the surface non-interface region. The enrichment analysis also showed that there were more neutral nsSNPs at the interface region than the buried region or the surface non-interface region of the protein structures. Further analysis of the neutral nsSNPs revealed that the majority of them came from cancerous samples. Later, David and Sternberg, 2015,

further characterised interface residues as 'core' and 'rim'. The authors argued that within interfaces, only a small subset of residues (hot-spots) is critical for the binding free energy of the protein-protein complex. They demonstrated that disease-causing mutations are preferentially located within the interface core, as opposed to the rim. In contrast, the interface rim was significantly enriched in neutral nsSNPs, similar to the remaining non-interacting surface. Additionally, David and Sternberg, 2015, found that energetic hot-spots tend to be enriched in disease-causing mutations, regardless of their occurrence in core or rim residues.

5.3 Methods

5.3.1 Analysis of CATH domain interactions

Protein domain interactions were analysed in the context of the CATH hierarchy for SCs and FunFams. Domain-domain interaction interfaces were defined as described in Section 2.3.5. In order to be able to analyse domain-domain interactions by interface type (or orientation) and within the structural perspective of the CATH hierarchy, the next sub-section describes the implementation of the iRMSD (interaction Root Mean Squared Deviation) method (Aloy et al., 2003; Jefferson et al., 2007b).

5.3.2 Analysis of domain-domain interactions by iRMSD

In order to classify interactions by their interaction interface, the relative orientation of the interacting pair was determined using the iRMSD method developed by Aloy

et al., 2003, and further developed by Jefferson et al., 2007b. This method determines if a pair of interacting domains bind in a similar orientation as another pair of interacting domains and thus if they are interacting using the same interfaces. The iRMSD method uses the sequence alignments generated by STAMP (described in Section 2.3.9), to match equivalent positions from each separate partner of the interaction pairs to determine the transformation of one structure to another.

As illustrated in Figure 2.3, comparing one pair of interacting domains AB to another pair $A'B'$, the transform of A' on to A and B' on to B is calculated by structure superimposition. The transform of A' to A is then used to transform A' to A and B' to B. Then for each domain of each pair, 7 representative coordinates are selected using the centre of gravity (centre of mass) as the middle point and then ± 5 Å in each axis from the centre of gravity to generate 6 further points. The RMSD of the seven representative points of A' ($7A'$) on to the 7 points of A ($7A$) and $7B'$ onto $7B$ is then calculated. The transform of B' on to B is then also used to transform A' on to A and the RMSDs of $7A'$ on to $7A$ and $7B'$ onto $7B$ calculated. The iRMSD for the A transform is the highest RMSD of the $7A'$ to $7A$ and $7B'$ to $7B$ using the A transform. The iRMSD for the B transform is the highest RMSD of the $7A'$ to $7A$ and $7B'$ to $7B$ using the B transform. The overall iRMSD is the lowest of the iRMSD using the A transform or the iRMSD using the B transform. In this work, the iRMSD method was modified so that the centre of gravity is determined based on the domain residues classified as structurally equivalent by STAMP (as defined in Section 3.3.1).

Aloy et al., 2003, suggested that interacting pairs with an iRMSD ≤ 5 Å should be considered similar whereas an iRMSD value as high as 5-10 Å could indicate similar

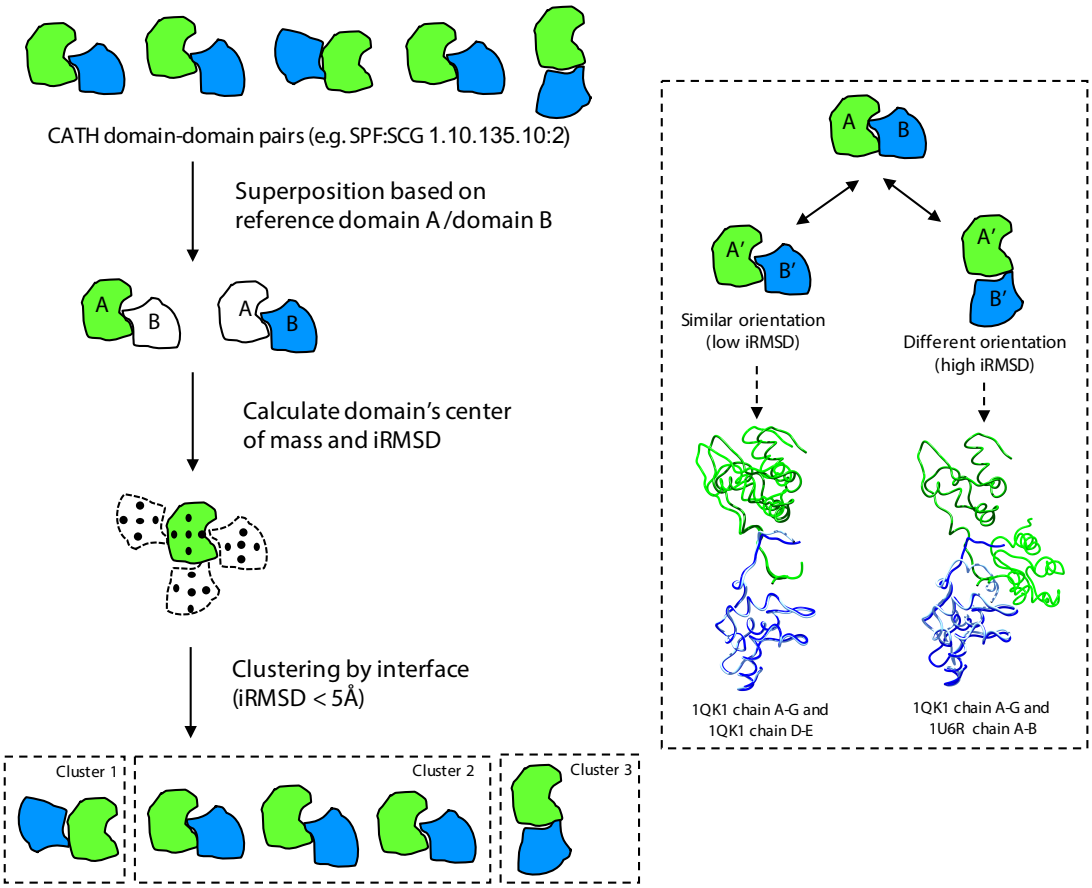


Figure 5.1: Schematic overview of the method for analysis of CATH domain-domain interactions by orientation. Domain-domain interactions are clustered by orientation based on a protocol that computes interaction RMSD (iRMSD) scores for related interacting domain-domain pairs. The example shown in the highlighted box illustrates the iRMSD analysis of inter-chain domain-domain pairs for PDB 1QK1 chain A-G (in green and blue) compared to PDB 1QK1 chain D-E and PDB 1U6R chain A-B (in light blue and light green), all belonging to CATH SPF 1.10.135.10 SC 2.

positioning of domains but with a rotation of one domain relative to another. In this work, the definition of a similar interaction is that two pairs of interacting domains are interacting with an iRMSD of <5 Å. Structure clustering based on the iRMSD score was performed by OC (Barton, G. J. University of Dundee, UK, 2004) using complete linkage. Percentage sequence identity (PID) was calculated according to the Doolittle method (Doolittle, 1981; Raghava and Barton, 2006), where PID = (identical positions / (aligned positions + internal gaps)), and ranges from 0 to 100.

5.3.3 Domain-domain contact propensities

Domain-domain interaction propensities were calculated based on a modified method by Glaser et al., 2001. The contact propensity for an interacting pair of residues i and j is given by:

$$CP_{ij} = \epsilon + \left(\frac{1}{\lambda}\right) \log\left(\frac{Q_{ij}}{(W_i/A_i)(W_j/A_j)}\right) \tag{5.1}$$

where $W_i = \frac{F_i}{N_r}$ corresponds to the normalised frequency of residue i. F_i is the number of residues i that have at least one contact with any residue across the interaction interface in the total number of interacting residues N_r. Similarly, the normalised frequency of contacts is $Q_{ij} = \frac{C_{ij}}{N_c}$, where C_{ij} is the number of contacts between residues i and j in the total number of observed contacts N_c. The expected number of contacts between residue i and j, is the value that would have been obtained if there were no preferences between residues of different types, i.e. W_i x W_j. This method was modified to account for the abundance of the amino acids in the human proteome by normalising the expected frequency of residue i (W_i) with the background abundance of residue A_i. Background relative amino acid frequencies for the human proteome were obtained from de Beer et al., 2013, and are summarised in Table 2.5. Arbitrary scaling factors $\epsilon = -1.5$ and $\lambda = 2$ were set such that the matrix contains easily comparable values.

As an illustrative example, the contact propensity for the interaction of Arg (i) and Glu (j) is obtained when considering: F_i = 73,490, F_j = 49,325 and N_r = 520,765, C_{ij} = 11,431 and $N_c = N_r/2$ = 260,383, which results in $W_i = 0.025$, $W_j = 0.013$ and $Q_{ij} = 0.044$. W_i is divided by 5.63%, which corresponds to the abundance

of Arg (A_i). Likewise, W_j is divided by 7.13% (A_j). The final contact propensity $CP_{ij} = 0.94$ (Figure 5.5) is obtained by applying Equation 5.1.

5.3.4 Domain-domain intermolecular bonding

Protein interactions were defined as described in Section 2.3.5. The types of intermolecular interactions analysed in this work include: hydrogen bonds; disulphide bonds; salt-bridges; aromatic-aromatic bonds; and non-covalent Van der Waals (hydrophobic) contacts. Intermolecular interaction types were calculated as defined in Section 2.3.5. Table 5.1 overviews which residue types participate in the various intermolecular bonds, noting that some only participate as donors or acceptors in the bounding. It is important to notice that the various interaction types defined were only probed on the basis of both residue-residue distance and residue-pairings (i.e. based on the identification of known residue/molecular-group acceptors and donors). The results of the analysis of domain-domain intermolecular interactions therefore used atomic distance and residue-pairing as proxies for intermolecular bonding classification.

5.3.5 Characterising genetic variants at domain interfaces

Genetic variants collected and organised in ProIntVar were separated into three main classes: germline variants; somatic variants; and disease-associated. Section 2.3.13 overviews the source and set of annotations collected for each variant. As described in Section 2.3.3 and Section 2.3.12, genetic variants were mapped to protein 3D structure, and domains were investigated in the context of functional families (FunFams) in CATH (Sillitoe et al., 2015). The analysis of genetic variants performed

Table 5.1: Summary of amino acid residues that participate in intermolecular interactions.

Bonding type	Contacting residues
Van der Waals[a]	Gly, Ala, Val, Leu, Ile, Pro, Phe, Tyr, Trp
Hydrogen bonds[b]	Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, Met, Set, Thr, Trp, Tyr
Salt-bridge[c]	Asp, Glu, Lys, Arg, His
Aromatic-aromatic[d]	Phe, Tyr, Trp
Disulphide bonds[e]	Cys

Intermolecular interaction types were calculated as defined in: [a]Van der Waals/Hydrophobic (Voet and Voet, 2010); [b]Hydrogen bonds (McDonald and Thornton, 1994); [c]Salt-bridges (ionic interactions) (Kumar and Nussinov, 2002); [d]Aromatic-aromatic (Burley and Petsko, 1985); and [e]Disulphide bonds (Schmidt et al., 2006).

in this Chapter focused on the subset of human missense variants described in Section 4.4.1 that map to STAMP-defined structurally conserved interaction interface residues (Section 2.3.7 and Section 4.3.2). Interaction interfaces were defined as: interaction with domains (*Inter. Domain*), interaction with ligands (*Inter. Ligand*, ligand definition is provided in Section 2.3.5), and interaction with other protein residues which are not part of CATH domains (*Inter. Protein*). The process of analysis of multiple genetic variants and interface residues, which mapped onto the same aligned positions (columns in the MSA), was performed as described in Section 4.3.2. Analyses of variation exchanges for amino acids and their physicochemical properties were performed as described in Section 4.3.4.

The terminology used to report nsSNPs is as follows: Cys34Phe represents an exchange from Cys residue in position 34 to Phe; Cys241* represents a mutation of Cys 241 to a stop codon (stop-gained mutation); and His661fs represents a frameshift

(fs) mutation, where His in position 661 is mutated in a way that leads to a change in the reading frame and the resulting translated product.

5.4 Results and discussion

5.4.1 Domain-domain interactions

The iRMSD measure is a useful solution to compare interacting domain pairs as it is independent of the size of the domains or interaction surface (Figure 5.1). The iRMSD method is also a purely geometric measure of interaction similarity. This is a useful property when the pair of interactions being compared are only remote homologs where equivalent residues can be difficult to define. In addition, an alternative measure, such as interface overlap, would not capture variations such as domain rotations and would not measure different degrees of identity. A potential problem with the iRMSD method for classifying interactions is that just because two pairs of domains are interacting at the same orientation does not necessarily mean that the interaction sites are the same.

Following the CATH hierarchy for SC and FunFam families under the Superfamily level (SPF), domain-domain interaction types can be classified as hetero-SPF or homo-SPF, based on the whether the two domains are grouped in different or the same CATH SPFs. Similarly, domain-domain interaction pairs can be classified as Hetero-SC/FunFam or Homo-SC/FunFam, for when the domains are grouped under the same or different families (SCs or FunFams in this case). Because SCs and FunFams are a family grouping that is defined under the Superfamily level in the CATH hierarchy, Hetero-SC/FunFam domain-domain pairs can arise from both

Hetero-SPF or Homo-SPF interaction types.

According to the CATH domain definitions and hierarchy, the structural dataset contains 90,132 unique domain-domain interaction pairs. Domain-domain interactions can be analysed by the context where the interaction occurs as: intra-chain, inter-chain or 'fusion'; if the domains are found in the same protein chain, in separate chains, or a mixture of both, respectively. Domain interactions were analysed for the CATH levels of SC and FunFam separately. As described in Section 2.3.5, two domains are considered to be interacting according to the distance between residues and if there are $\geq$10 interacting residue pairs between the domains. Results on the unique domain-domain interaction types are summarised in Table 5.2.

Domain-domain interactions were further classified by orientation based on an improved interaction RMSD (iRMSD) method (Aloy et al., 2003; Jefferson et al., 2007b). The iRMSD protocol requires at least two domain-domain entries of the same domain-domain interaction type (each domain belonging to a particular SPF-SC/FunFam), because a single entry would result in no possible comparisons. A minimum of two entries results in only one possible interaction iRMSD cluster (or interaction orientation). Since the iRMSD protocol relies on the multiple structure alignment of the involved domains, SCs and FunFams containing a single domain member were not considered. The domain-domain interaction analysis was performed for the three main types of interaction: inter-chain, intra-chain and fusion as shown in Table 5.2.

Table 5.2: Overview of CATH domain-domain interaction analysis by iRMSD. Results are shown for both structural clusters (SCs) and functional families (FunFams). The number of unique domain-domain interaction types for inter-chain, intra-chain and fusion, as well as comparing the CATH hierarchy classification (for Superfamily and SC/FunFam) of each domain-domain pair is shown.

SC domain-domain types	**Inter-chain**	**Intra-chain**	**Fusion**
All	1,837	1,029	2,604
Homo-SPF	826	140	905
Hetero-SPF	372	594	873
Homo-SC	795	52	808
Hetero-SC	403	682	970
Hetero-SC (Homo-SPF)	31	88	97
Hetero-SC (Hetero-SPF)	372	594	873
SPF-SPF	893	585	1,331
SC-SC	1,198	734	1,778

FunFam domain-domain types	**Inter-chain**	**Intra-chain**	**Fusion**
All	3,345	1,778	4,748
Homo-SPF	1,338	300	1,583
Hetero-SPF	522	992	1,373
Homo-FunFam	1,201	13	1,208
Hetero-FunFam	659	1,279	1,748
Hetero-FunFam (Homo-SPF)	137	287	375
Hetero-FunFam (Hetero-SPF)	522	992	1,373
SPF-SPF	892	594	1,335
FunFam-FunFam	1,860	1,292	2,956

5.4.2 Classifying domain-domain interactions by orientation

The next step in the analysis of domain-domain interactions is the classification of the type of interaction by orientation with the iRMSD clustering protocol. Domain-domain interaction pairs were compared all-against-all where the domain-domain interaction was of the same type (i.e. both interacting domain sets sharing the

same CATH classification) and an iRMSD score computed for each pair. Complete linkage hierarchical clustering based on the iRMSD score was then performed using an inclusion threshold of $\leq$5 Å.

The distribution of the number of clusters (or modes of interaction) resulting from the domain-domain classification by iRMSD is shown in Figure 5.2. The trend in the distribution of the number of clusters is similar for SCs and FunFams. Nevertheless, the overall number of clusters inclusion is lower for FunFams when compared to SCs. The difference between FunFams and SCs results from the fact that although the total number of CATH FunFams is higher than that of SCs, the number of domains per FunFam group is lower than the observed for SCs. In this way, more structural diversity is expected for SCs. The number of clusters is lower for intra-chain domain-domain interaction types when compared with both inter-chain and

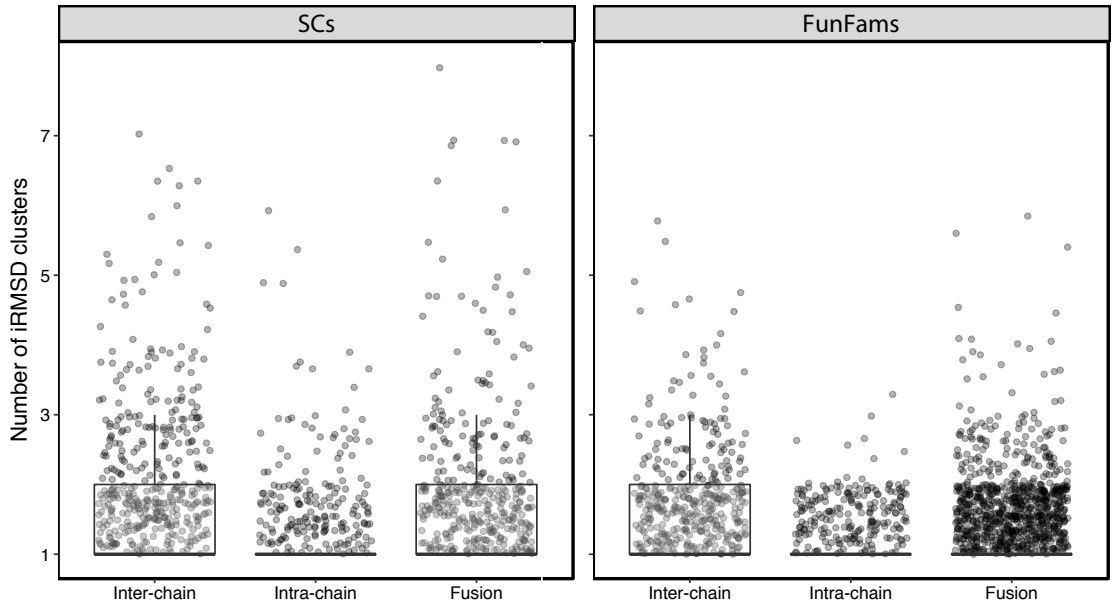


Figure 5.2: Box-plots showing the distribution of the number of iRMSD clusters for SCs and FunFams. The iRMSD scores were obtained for both SC and FunFam groups under Superfamily level in CATH, and the various domain-domain interaction types: inter-chain, intra-chain and fusion.

fusion interaction modes, for both SCs and FunFams. This result is in agreement with the observation that multi-domain proteins share a more structural constrained fold/topology (Bhaskara and Srinivasan, 2011; Vogel et al., 2004).

Figure 5.3 shows the relationship between iRMSD score and protein sequence identity (PID), as originally plotted by Aloy et al., 2003. The original analysis was performed using a much smaller and non-redundant set of interacting protein domains up to the Fold level as classified in SCOP (Lo Conte et al., 2000), where only the best iRMSD scoring domain-domain entry for each domain-domain interaction type was plotted. In contrast, here the iRMSD classification was performed for SCs and FunFams under Superfamily level in CATH, which groups together very structurally similar domains, thus an inherently high level of sequence redundancy was kept. Although not fully comparable, the results shown here seem to be in overall agreement with the original study. Interestingly, as suggested by the high density of points on the right-hand side of the plots (Figure 5.3 A and B), a higher number of domain-domain pairs interact using distinct interfaces, although sharing a high PID. There are also some differences in the SCs and FunFams profiles, which result from the different domain composition observed for these two CATH Superfamily sub-groups. An additional aspect expected to introduce some confounding interference is related to the way in which PID is calculated (Raghava and Barton, 2006). Aloy and colleagues calculated PID as the number of identical residues divided by the number of structurally equivalent residues, whereas here the Doolittle method (Doolittle, 1981), which also considers internal gap positions, was used. PID ranges from 1.8% to 100% with the mean PID being 63%.

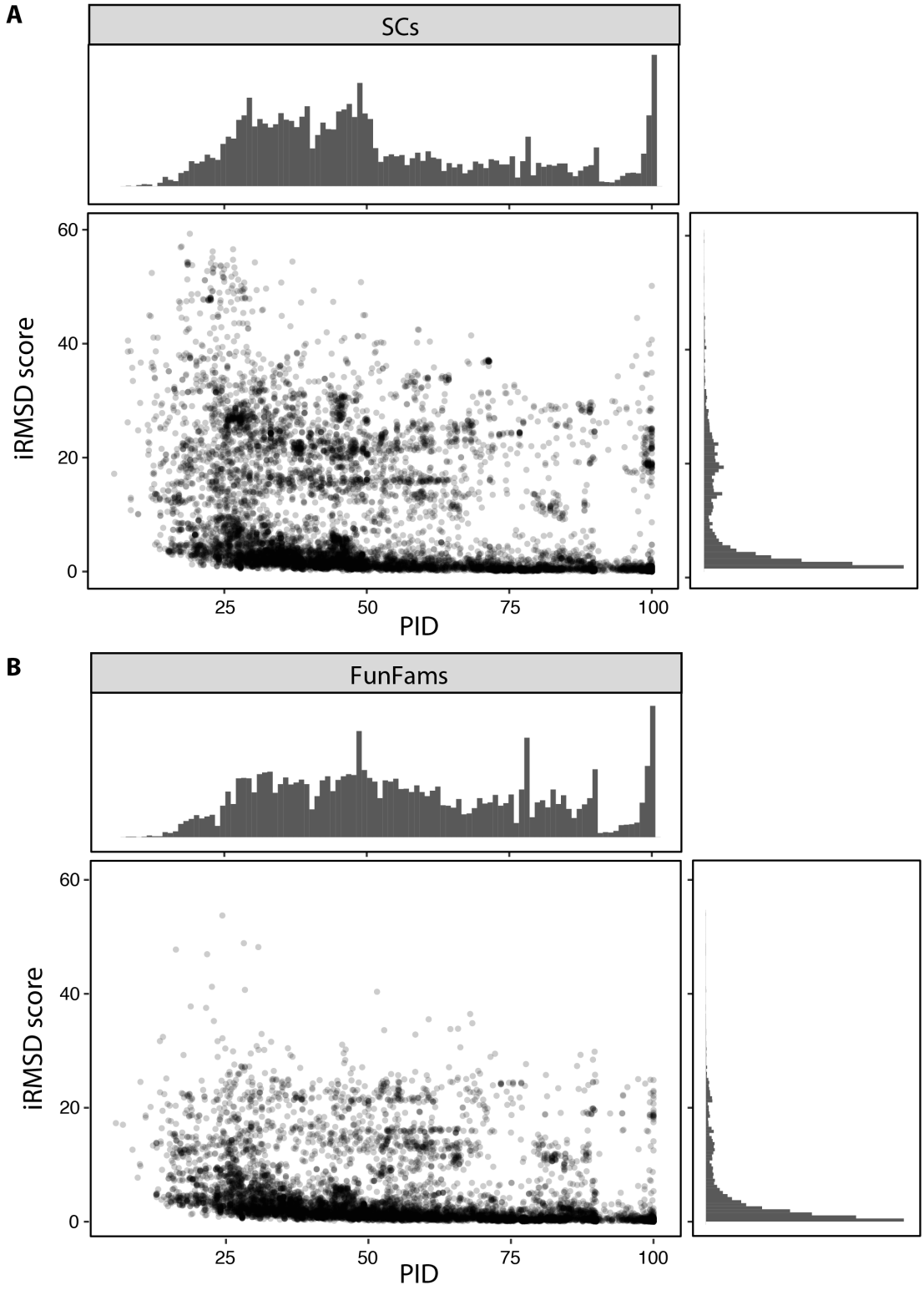


Figure 5.3: Scatter plot showing the correlation between interaction RMSD (iRMSD) *versus* percentage sequence identity (PID). The iRMSD PID scores were obtained for both: A) structural clusters (SCs) and B) functional families (FunFams) under Superfamily level in CATH. Accompanying histograms highlighting the binned counts per PID and iRMSD are also shown.

5.4.3 Mapping genetic variants to conserved sites at the interfaces

Initial analysis of genetic variants mapped to protein domain interfaces was performed in Chapter 4. In order to extend the analysis of variation at domain interfaces, CATH FunFam protein domain families were explored by considering STAMP-defined structurally conserved regions/positions (SCRs) that are found to participate in interactions with: domains (domain-domain interaction); ligands (domain-ligand interaction); and protein (domain-protein, where protein correspond to protein residues not classified to any CATH domain). Variation analysis was performed for the subset of human missense variants described in Section 4.4.1.

Structurally and functionally important residues are often well conserved (Hu et al., 2000; Guharoy and Chakrabarti, 2010; Fiser et al., 1996; Thusberg et al., 2011). It is reasonable to assume that key residues of a vital protein-protein interaction are spared from mutations, as the loss of an interaction either has fatal consequences for the organism or confers an evolutionary disadvantage. Consequently, functional interfaces are commonly associated with a lower mutability than other, non-functional parts of the protein surface.

Table 5.3 shows the total number of FunFams for which variants could be mapped to their conserved sites within the domain interaction interfaces. The number of FunFam families drops to 1,080, 915, 202 when considering germline, somatic and disease-associated variants only, respectively, when compared to the 1,127 families when accounting for all variants (Table 4.1). The total number of unique germline variants that map onto interfaces is 9,574. This accounts for a mean of 11 variants

Table 5.3: Overview of FunFams for which genetic variation could be mapped to structurally conserved residues at domain interaction interfaces. Genetic variants were grouped as germline variants, somatic mutations and disease-associated variants. The mean ($\bar{x}$) [minimum; median; maximum] and the total number of genetic variants that were mapped to domains in FunFams is shown.

	FunFams		
Variants	**Total FunFams**	**Variants per FunFam ($\bar{x}$)**	**Total variants**
All	1,127	17 [1; 10; 1053]	16,005
Germline[a]	1,080	11 [1; 6; 395]	9,574
Somatic[a]	915	7 [1; 4; 481]	4,934
Disease[a]	262	7 [1; 3; 177]	1,497

[a]Final subset of human missense mutations mapped onto STAMP structurally conserved positions within interaction interfaces.

per FunFam. The total number and mean of variants is lower for the somatic and disease-associated classes, with a total 4,934 and 1,497 variants (mean of 7 residues), respectively. This is expected since only 15% of all variants map to interface positions in FunFams (Figure 4.5), and a decreasing number of germline (64%), somatic (30%) and disease-associated (6%) variants in each class are analysed (Table 4.1).

Figure 5.4 summarises the total number of variants mapped to STAMP conserved sites across the interaction interfaces. The breakdown of the number of variants that map onto interfaces by interface type reveals that a majority of variants (73-78%) is involved in domain-domain interactions. The remaining 13-18% and 0.9-1.1% of the variants are mapped to domain-ligand and domain-protein interactions, respectively. These results are in line with the proportions shown in Figure 4.5, which were calculated taking into account both conserved and variable positions within interaction interfaces. This results from the fact that a high number of interface sites with mapped variants correspond to conserved positions. In fact, the number of STAMP-defined structurally conserved regions account for the mapping

of 82-86% of all variants (Figure 5.4).

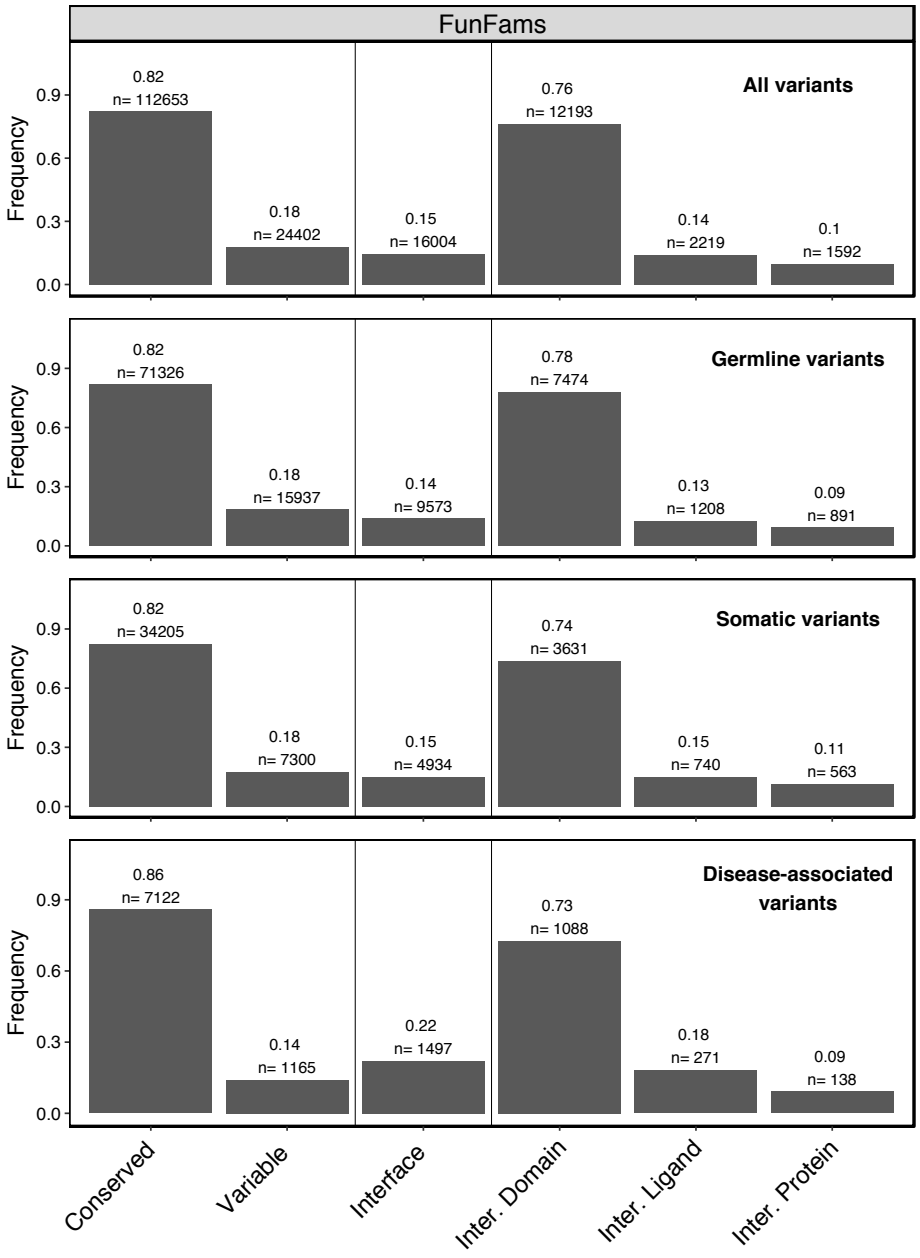


Figure 5.4: Distribution of genetic variants mapped to structurally conserved regions within interaction interfaces in FunFam domain families. Genetic variants were grouped as germline variants, somatic mutations and disease-associated variants. The aggregation of all variants in the three classes is also provided. n corresponds to the total number of genetic variants in each class that fall into that particular structure environment. Structural environments/regions were defined as described in Section 4.3.2. The number of variants in variable regions is shown for comparison purposes.

5.4.4 Domain-domain interaction propensities

In order to investigate whether some amino acid residues are preferentially mutated at interfaces, domain-domain contact propensities were derived from the residue pairings observed for the interaction interfaces of the subset of 1,127 FunFam families, to which variation could be mapped (Table 5.3). Figure 5.5 shows the log contact propensity matrix for all amino interaction pairs that compose the domain-domain interfaces. The contact propensity matrix is symmetric since the amino acid pairing frequencies are non-directional.

Among the most frequent amino acid pairs are Arg-Glu, Arg-Asp, Lys-Glu, and Lys-Asp, which correspond to favourable opposite-charge ionic interactions (salt-bridges). Leu-Leu interaction is also enriched and results in favourable hydrophobic (aliphatic-aliphatic) interactions. Other hydrophobic (Van der Waals) interactions, as well as hydrogen bonding, are favourable including Leu-Val, Val-Val, Ser-Ser and Ile-Ile. Aromatic-aromatic interactions between Phe-Phe and polar interactions between Ser-Asp are also observed. The contribution of Asp to the maintenance of interaction interfaces might be related to its special conservation status (Fiser et al., 1996). Cys-Cys interactions are also enriched, which might result from favourable disulphide bonds. Among the less likely interactions are interactions between large aromatic and hydrophobic residues (Trp, Tyr, Ile) and sulphur-containing residues (Cys and Met). These correspond to chemically complex amino acids which are less favourable at domain-domain interfaces. Interactions with Trp are expectedly less likely since Trp is the least abundant amino acid in the proteomes (Table 2.5). The interaction between Trp-Cys is also very unfavourable, given the nature of

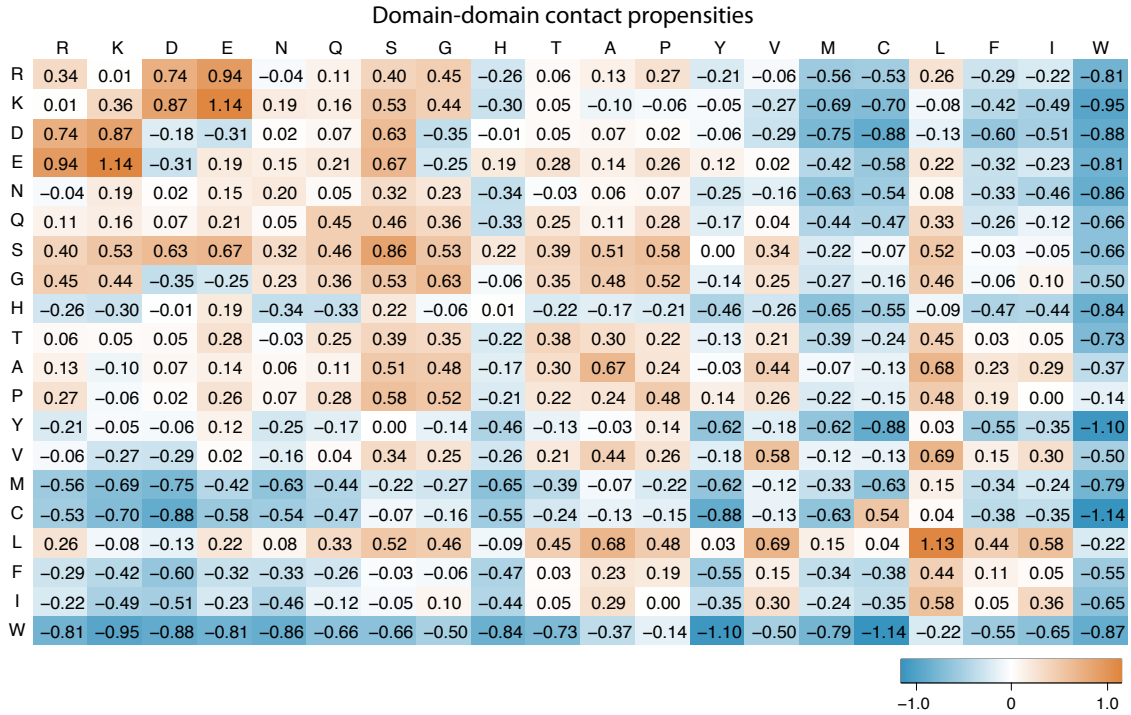


Figure 5.5: Log contact propensities for variant-mapping FunFam protein families. The amino acid pairings are provided for domain-domain interactions observed for FunFam protein families for which variants could be mapped to conserved positions. Amino acids are arranged by one letter code according to increasing hydrophobicity (least hydrophobic is left/top and most hydrophobic is right/bottom) according to the Fauchère et al., 1988, scale.

the physicochemical properties of these residues and their abundance. The least favourable interactions are observed between some of the biggest amino acid residues including Arg, Lys, Asp and Glu to aromatic residues, Trp and Phe. Accordingly, charged-aromatic interactions including: Trp-Arg, Trp-Lys, Trp-Glu, Trp-Asp and Phe-Asp; are depleted. This is expected since aromatic residues share constrained side-chain geometries that impose limitations on the angles which would allow for electrostatic interactions to be made with these residues. Overall, these results are in agreement with work by Glaser et al., 2001, which calculated protein-protein contact propensities using a non-redundant set of protein complexes.

Figure 5.6 shows the comparison of the abundance of amino acids in the human

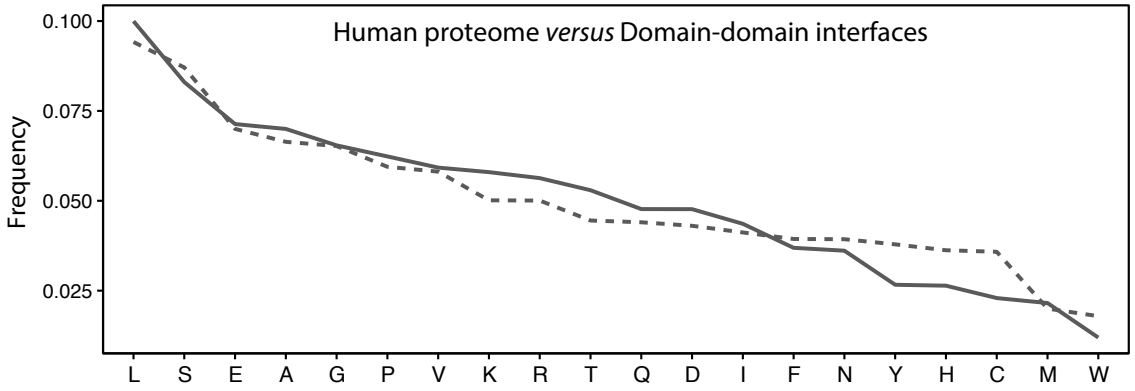


Figure 5.6: Comparison of the abundance of amino acids in the human proteome and in the domain-domain interaction dataset. Amino acid frequencies are provided for the human proteome (line) and the domain-domain interface subset (dashed-line). Amino acids are arranged by 1-letter code according to decreasing abundance in the human proteome (de Beer et al., 2013).

proteome (de Beer et al., 2013) and the abundance of amino acids in the domain-domain interfaces dataset. The frequency of each amino acid in the domain-domain interface dataset follows a similar overall trend to the frequency of each amino acid in the human proteome. A significant depletion of charged residues (Lys, Arg and Asp), as well as aromatic Thr, is observed. In contrast, an enrichment of Tyr, His and Cys is observed at the domain-domain interfaces. This result indicates that despite the natural abundance of each amino acid, particular residues are preferred at interfaces. It also agrees with the contact propensities observed in Figure 5.5 and those reported by Glaser et al., 2001.

Table 5.4 overviews the occurrence of particular intermolecular bonding types observed in the domain-domain interaction interfaces dataset. Non-covalent Van der Waals (hydrophobic) intermolecular interactions account for 68% of all interactions in the dataset. Hydrogen bonds are formed for 15% of the amino acid pairs, followed by salt-bridge (ionic) interactions. The least common interaction types are aromatic-aromatic (1%) and disulphide bonds (0.001%).

Table 5.4: Overview of the domain-domain interaction bonding types in the FunFam protein families.

Bonding type	**Total number of contacts**
Van der Waals	541,727 (0.68)
Hydrogen bonds	122,959 (0.15)
Salt-bridge	57,962 (0.07)
Aromatic-aromatic	11,655 (0.01)
Disulphide bonds	962 (0.00)

5.4.5 Analysis of genetic variation amino acid exchanges at interfaces

With the growing amount of protein-protein interaction information that exists today, studying the effects of variation on interfaces is essential to understanding the molecular mechanisms of disease. In a similar fashion to the analysis performed in Chapter 4, genetic variation amino acid exchanges were investigated for structurally conserved positions at the interfaces. The overall trend in amino acid exchange frequencies is similar when comparing different variation classes as well as when comparing the interface subset with the entire variation dataset, as performed in Section 4.4.4.

Figure 5.7 shows the frequency difference obtained for 'from' and 'to' amino acid exchanges by comparing the relative frequencies of the entire variation dataset and those of the interface subset. The frequency difference is calculated as the frequency of amino acid exchange at the interface, minus the frequency of exchange in the entire variation dataset. Frequency differences bigger than zero correspond to enrichment of a particular amino acid exchange at the interfaces, whereas frequency differences lower than zero correspond to interface depleted variation exchanges. Only minor

frequency differences are observed overall. Nevertheless, the difference is bigger for 'from' amino acid exchanges, than for 'to' exchanges.

Regarding 'from' exchanges, Arg mutation is enriched at interfaces (p-value $<$ 0.01), whereas Ser and Gly are depleted, for all three variation classes (Figure 5.7). Mutation from these to small size amino acids is likely to introduce steric clashes. Likewise, mutation from Arg is likely to affect binding by introducing a void space in the structure, which might affect interaction (Martin et al., 2002; Stefl et al., 2013). There is an enrichment of mutation of Asp at interfaces for disease-associated variants, which potentially results from the disruption of hydrogen bond networks. Mutation from Cys is enriched for somatic variants. This mutation is likely to affect protein binding, particularly to ligands, since Cys participates in disulphide bonds and is often found coordinating binding to metal ions.

For 'to' amino acid exchanges minor frequency differences are observed for amide-containing Asn and Gln residues (Figure 5.7), which are found enriched (p-value $<$ 0.05) at interfaces and are major contributors as donors and acceptors in hydrogen bonds. In contrast, neutral residues (Ser and Leu) are depleted at interfaces, particularly for known disease-associated variants. Overall, mutations are expected to affect the interaction interfaces to some extend, through several mechanisms such as by introducing constrained geometries to the protein backbone (Pro), introducing steric clashes (Arg) or void spaces (Gly), and disrupting interaction networks (Asn and Phe).

Similarly to the breakdown of amino acid exchanges obtained for the entire variation dataset (Figure 4.9), the breakdown of all amino acid exchanges observed in the different genetic variation subsets that map onto STAMP-defined structurally

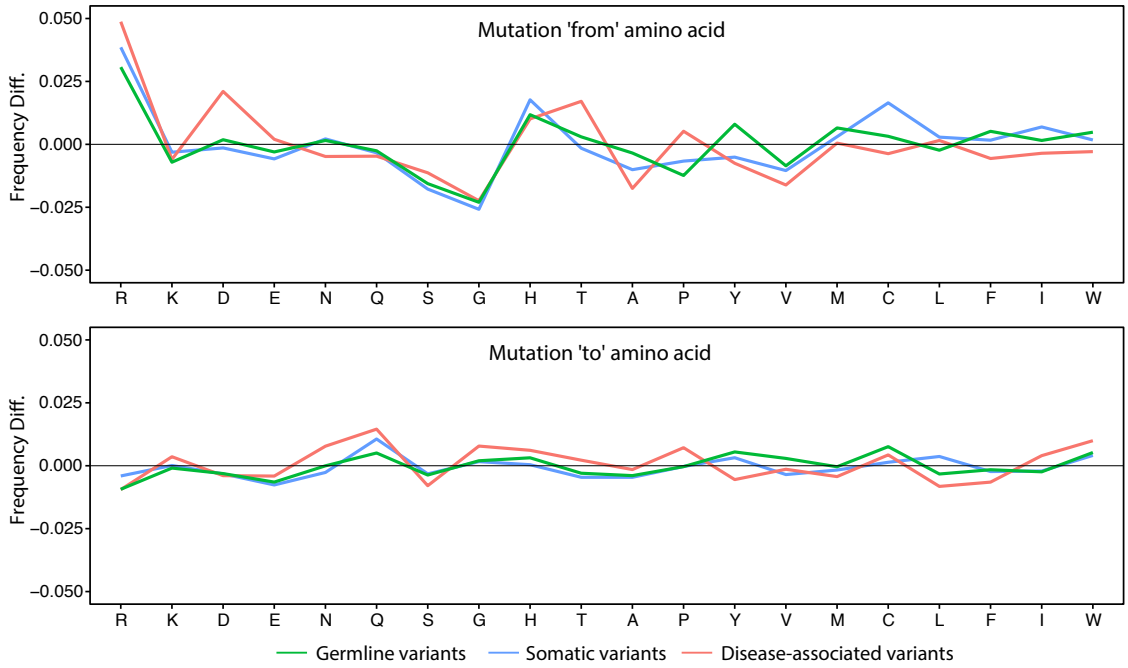


Figure 5.7: Comparison of the mutation frequency differences for all classes of genetic variants. The frequency difference of observed mutations is provided for germline variants (green), somatic mutations in (blue), and disease-associated variants (red). Amino acids are arranged by 1-letter code according to increasing hydrophobicity (least hydrophobic is left and most hydrophobic is right) according to the Fauchère et al., 1988, scale.

conserved interface residues is provided. An identical amino acid exchange profile is observed when comparing the two variation datasets. Figure 5.8 shows the frequency difference obtained between the interface variation subset and the overall variation dataset. Bigger frequency differences are observed for disease-associated amino acid exchanges mutating from hydrophobic and polar residues (Asp, Gln, Ser, Ala, Pro, Tyr, Val, Leu, Phe and Ile). Mutations to Pro are not particularly enriched (p-value > 0.05) at interfaces, despite being among the most frequent mutation found for disease-associated variants (Figure 5.7).

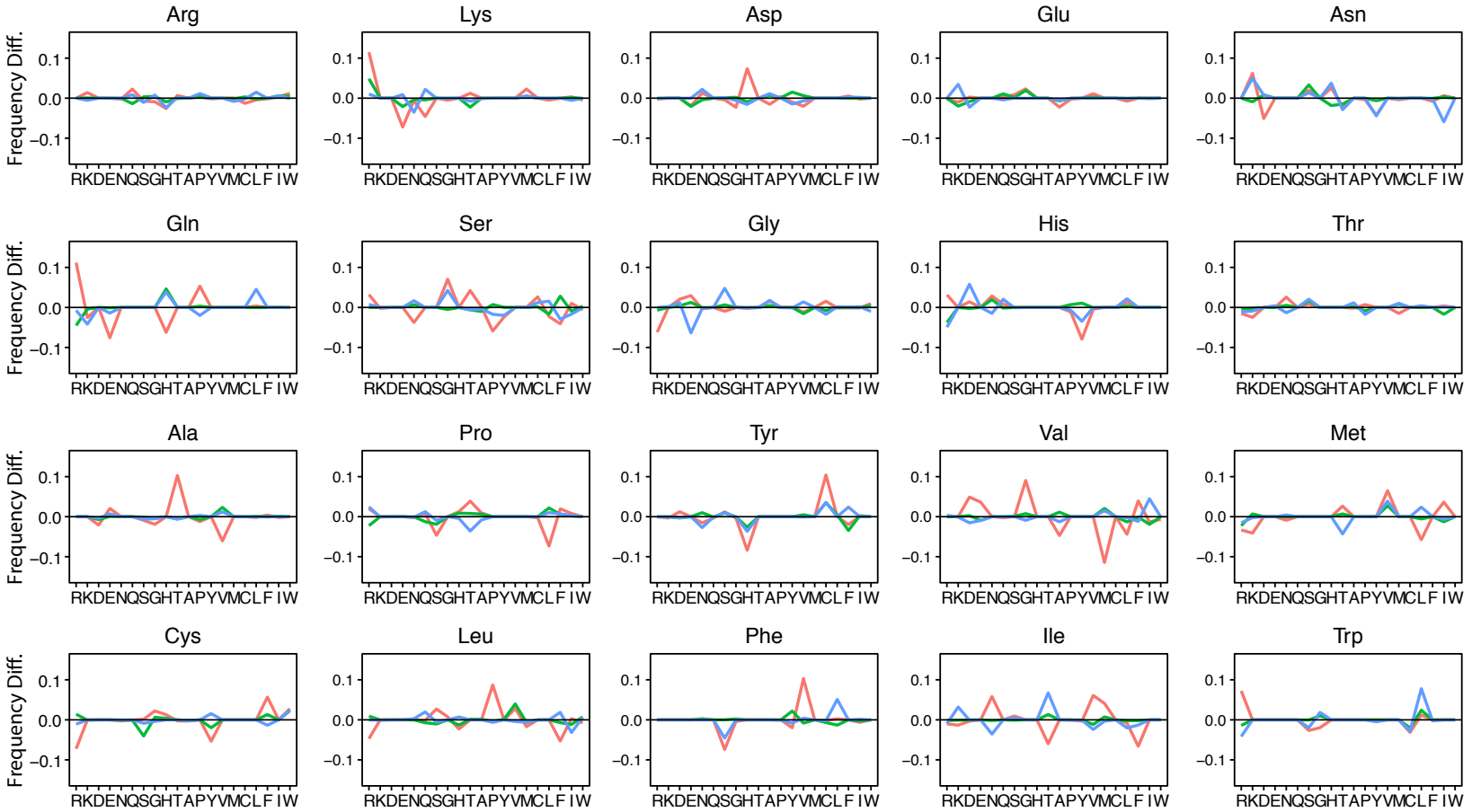


Figure 5.8: Comparison of the mutation frequency differences obtained for all classes of genetic variants, when comparing conserved interface residues and against the whole database. The frequency difference of observed mutations is provided for germline variants (green), somatic mutations in (blue), and disease-associated variants (red). Each plot shows the frequency difference of mutation from a specific amino acid (e.g. Arg at top left) to every other amino acid. Amino acids are arranged by 1-letter code according to increasing hydrophobicity (least hydrophobic is left and most hydrophobic is right) according to the Fauchère et al., 1988, scale.

Amino acid exchanges involving residues that participate in favourable domain-domain interfaces (Figure 5.5) are enriched for disease-associated variants, since they lead to drastic changes (loss of interaction), based on the properties of the amino acids. An example of such transitions would be for opposite charge mutations from Lys to Glu and Asp to His (depleted outside interfaces). These lead to opposite charge transitions and potentially disrupt ionic interactions as well as hydrogen-bond networks. The scenario in which amino acid exchanges involving residues that participate in unfavourable (gain of interaction) interactions lead to a more favourable interaction is also observed. Mutations from Val to Met, Tyr to His, Met to Leu, and Ile to Tyr, are among those transitions which are depleted for disease-associated variants and might lead to improvement of the interaction. Disease-associated variant exchanges from Ala to Thr are highly enriched (p-value < 0.001) at interfaces, despite both amino acids sharing similar properties. Only minor frequency differences are observed for the mutational profiles obtained for germline and somatic variants, when compared to those from disease-associated variants. Mutation of Gly to Glu is depleted at interfaces for somatic variants.

5.4.6 Analysis of genetic variation exchanges at interfaces according to physicochemical properties

In order to further investigate specific differences observed for amino acid exchanges at interfaces, amino acids were grouped according to their physicochemical properties. Similarly to the analysis performed in Section 4.4.5, two single state alphabets were investigated.

Figure 5.9 shows the frequency difference obtained between the interface variation subset and the entire variation dataset, when considering the physicochemical alphabets. Again, only minor differences are observed for the 'from' and 'to' physicochemical group exchanges across the three variation classes. The analysis of exchanges for both physicochemical alphabets reveals that only minor frequency differences are observed when comparing variants in different variation classes, as well as when comparing to the frequency profiles obtained for the full variation dataset (Figure 4.11).

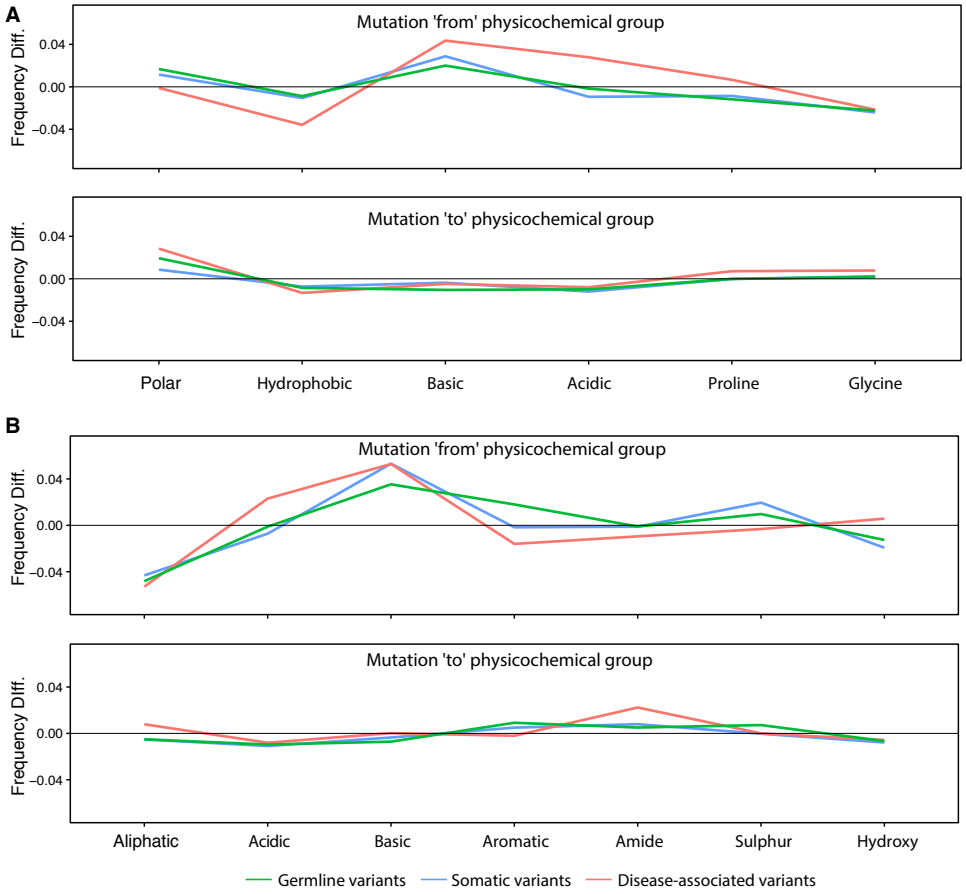


Figure 5.9: Frequency differences for physicochemical exchanges observed for the three classes of genetic variants obtained when comparing conserved interface residues and all domain residues. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red). Amino acids were grouped into two alphabets: A) 'Chemical_A', and B) 'Chemical_B'.

In agreement with the amino acid exchange profiles (Figure 5.8), mutations from Gly and mutations to Pro are slightly enriched (p-value > 0.05) for disease-associated variants. A slight enrichment (p-value < 0.05) in the frequency of mutation from both acidic (Arg, Lys and His) as well as basic (Asp and Glu) residues is observed at interfaces, which is accompanied by a depletion of hydrophobic residues (Figure 5.9 A). This is particularly noted for disease-associated variants. Mutation to amide-group containing residues (Asn and Gln) also are slightly enriched at interfaces for disease-associated variants, although this enrichment is not significant (p-value $>$ 0.05) (Figure 5.9 B).

Figure 5.10 shows the breakdown of the frequency differences for physicochemical changes introduced by genetic variants at interfaces according to the two alphabets defined. To help to identify particular physicochemical trends at the interfaces, Figure 5.10 shows the frequency difference obtained between the interface variation subset and the entire variation dataset. Regarding the Chemical_A alphabet, the most dramatic transition is observed for exchanges between hydrophobic and polar residues, as well as Pro and Gly, for disease-associated variants. Smaller changes in the frequency of exchanges between basic residues is observed for somatic variants (Figure 5.10 A). Mutations of hydrophobic residues to Pro residues observed for disease-associated variants are statistically significant (p-value < 0.01). Enrichment of exchanges from basic and Cys/Met residues to aliphatic is accompanied by a depletion of basic and hydrophobic residues, for disease-associated variants (Figure 5.10).

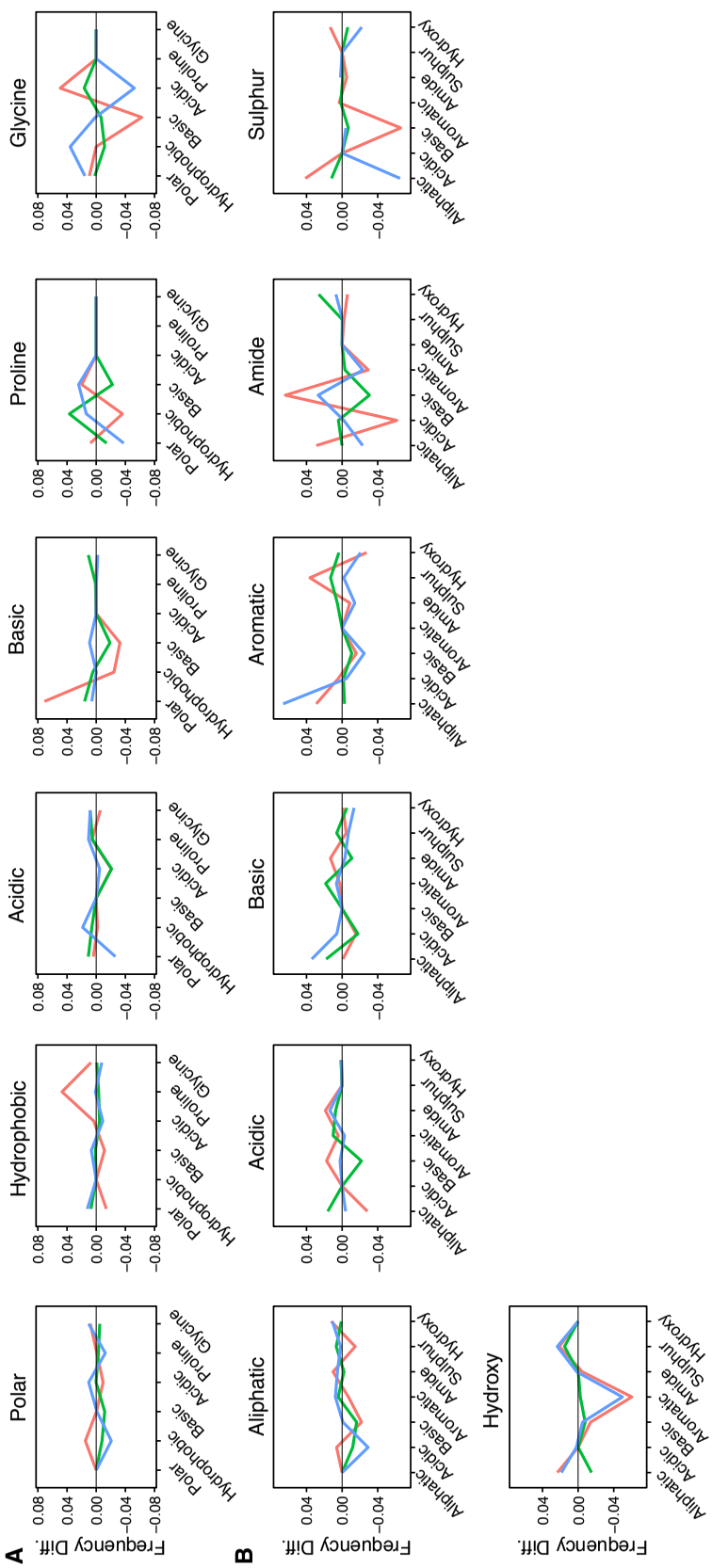


Figure 5.10: Comparison of the physicochemical exchange frequency differences for all classes of genetic variants. Genetic variants were grouped as germline variants (green), somatic mutations (blue) and disease-associated variants (red). Amino acids were grouped into two alphabets: A) 'Chemical_A', and B) 'Chemical_B'.

Similarly to what was observed for the entire variation dataset (Figure 4.12), the analysis of physicochemical transitions shown for the Chemical_B alphabet indicate that mutation of hydroxy-containing residues to aromatic for both somatic and disease-associated variants is enriched at interfaces. Similarly, mutation of residues containing sulphur (Met and Cys) also show depletion in exchanges to aliphatic and basic residues, for somatic and disease-associated variants, respectively (Figure 5.10 B). In fact, the bigger frequency differences observed between somatic variants and germline variants are observed for exchanges from Met and Cys residues (sulphur group) to chemically complex aliphatic residues, as well as aromatic residues in the case of somatic variants. Transitions from residues that contain hydroxy-groups (Ser, Thr) to aromatic residues is depleted for somatic and disease-associated variants at the interfaces (Figure 5.10). Amide-group containing residues (Asn and Gln) display an enrichment (p-value < 0.05) of exchanges to basic residues, which is accompanied by a depletion of acidic residues, for disease-associated variants. Among the biggest frequency differences observed at the interfaces for somatic variants when comparing to germline variants are also observed for Gly to hydrophobic (enrichment) and from Gly to acidic (depletion). Enrichment of aromatic to aliphatic residues is observed for somatic variants at the interface (p-value < 0.05). Among the more dramatic depletions of somatic variants are the observed depletion of Met/Cys to aliphatic, and Ser/Thr to aromatic residues. The frequency differences obtained for germline variants are the smallest overall, which results from the neutral nature of this variation.

5.4.7 Prioritising the analysis of genetic variants at interfaces within FunFam families

The residue variation at key sites within a protein may result in a series of changes impacting protein stability, the conformation or folding kinetics, the disruption of salt bridges and hydrogen bonds, and the perturbation of the energy landscape, ultimately affecting protein function (Wang and Moult, 2001; Yue and Moult, 2006; Teng et al., 2010; Stefl et al., 2013). The comparison of the variation profiles in terms of amino acid exchanges and physicochemical properties obtained for different classes of variants enhances features that can be used to help prioritise the analysis of the potential consequences of variants. In particular, trends observed for disease-associated variants at the interfaces and across all structural environments can be used to help investigate germline variants thought to be neutral.

Table 5.5 summarises the general features used to select FunFam protein families and domains for further analysis. Variation consequences have been grouped according to the main consequence or behaviour they can have an effect on (Wang and Moult, 2001; Martin et al., 2002; Tang et al., 2004; Stone and Sidow, 2005; Khan and Vihinen, 2007). Variation can lead to: disruption of intermolecular bonds (e.g. hydrogen and disulphide bonds); substantial size changes (e.g. introduction of steric clashes or void spaces); change in ionisation state (e.g. change to opposite charge); change in hydrophobicity; and mutation from/to Pro and Gly. Another level of filtering can be performed by focusing on variants predicted to be deleterious/damaging (Figure 4.15).

A set of FunFam protein families were selected as proof-of-concept examples to

Table 5.5: Summary of the most important variation consequences used to help to prioritise the analysis of genetic variation.

Feature	**Evidence**
Disruption of intermolecular bonds	Enrichment of mutations that lead to drastic physicochemical transitions found for disease-associated variants (e.g. Figure 5.8 and Figure 5.10).
Substantial size changes	Broader distribution of volume and atomic mass transitions for disease-associated variants, as well as enrichment of mutations that lead to size changes (e.g. Figure 4.13 and Figure 5.8).
Change in ionisation state	Enrichment of opposite charge exchanges found for disease-associated variants (e.g. Figure 5.8 and Figure 5.10).
Change in hydrophobicity	Broader distribution of hydrophobicity transitions for disease-associated variants (e.g. Figure 4.13 and Figure 5.8)
Mutation from/to Pro and Gly	Enrichment of exchanges to Pro and mutations from Gly, for disease-associated variants (e.g. Figure 5.8 and Figure 5.10).

showcase the power of performing structure-based analysis of genetic variation in the context of feature-rich annotations and within protein families. Protein families were selected if variants were mapped to structurally conserved domain-domain and/or domain-ligand interfaces and at least one of the variants displays some of the features defined in Table 5.5. Structurally conserved MSA positions at the interface where the variation exchange outcome is observed in that particular column (Table 4.5) were excluded, as they are less likely to impair activity/stability of the interface. Variation types other than missense variants (i.e. frameshift-variant, stop-gained, and others described in Section 4.3.1), were also included in the proof-of-concept analysis. Each potential consequence shown in Table 5.5 was scored according to a decreasing scale, from 6 to 1, and every FunFam family showing mapping of genetic variants at interfaces was ranked. The score of each potential consequence was added

to a particular FunFam, when at least one instance of that consequence is observed for that FunFam. Figure 5.11 shows the clustering analysis results performed on the 40 highest ranked FunFam families.

Table 5.6 lists a set of 20 highly-ranked FunFam protein families and domain examples. The examples include a variety of enzyme families: proteinases; dehydrogenases; esterases; reductases; and isomerases; as well as regulatory protein families;

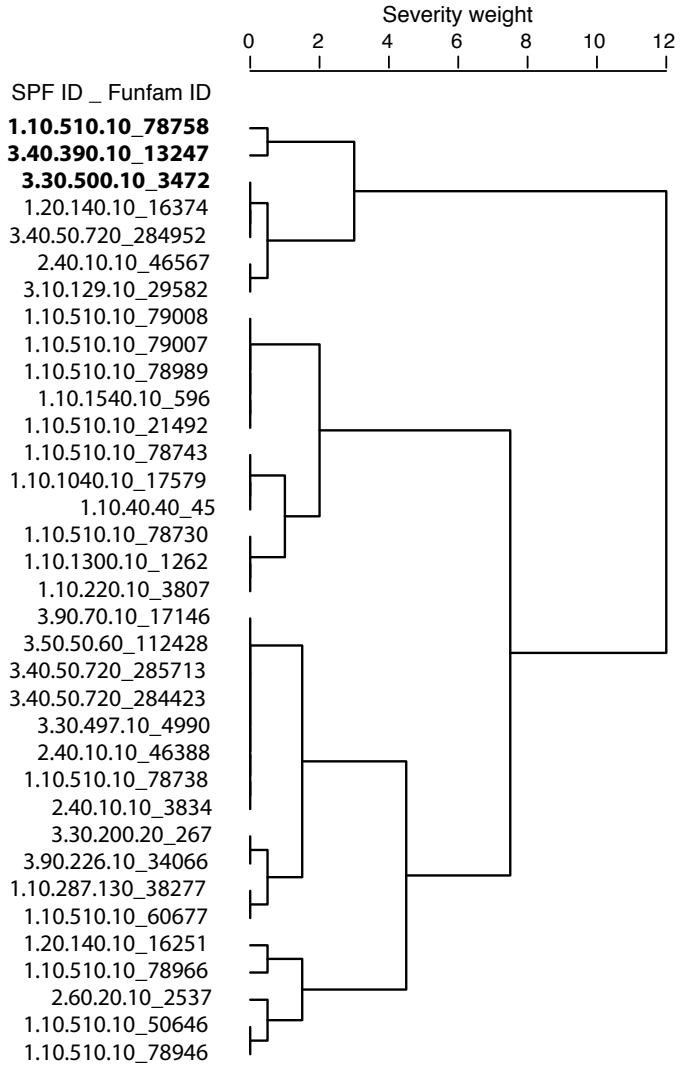


Figure 5.11: Clustering analysis of the top 40 FunFam families mapping potentially disruptive genetic variants at interfaces. The potential consequences of variation summarised in Table 5.5 were used to score each FunFam family (severity weight). Clustering was performed with *hclust* method in R (R Core Development Team, 2016) using Euclidean distance and complete linkage.

signalling receptors and immunity proteins.

Table 5.6: Top 20 FunFam families showing potentially disruptive genetic variants at the interfaces.

SPF ID	FunFam ID	FunFam Descritption/Representative domain
1.10.510.10	78758	Ephrin type-A receptor 2 (ephrin-A2)
3.40.390.10	13247	Matrix metalloproteinase-3 (MMP-3)
3.30.500.10	3472	T-cell surface glycoprotein CD1e (CD1e)
1.20.140.10	16374	Acyl-Coenzyme A dehydrogenase family, member 9
3.40.50.720	284952	Estradiol 17-β-dehydrogenase 1
2.40.10.10	46567	Coagulation factor XI (FXI)
3.10.129.10	29582	Acyl-coenzyme A thioesterase 12
2.60.20.10	2537	β-crystallin A3
3.30.200.20	267	Fibroblast growth factor receptor 1 (FGFR-1)
3.30.497.10	4990	α-1-antitrypsin 1-3
3.50.50.60	112428	Thioredoxin reductase 1
3.90.226.10	34066	Enoyl-CoA delta isomerase 2, mitochondrial
3.90.70.10	17146	Cathepsin B
1.10.1040.10	17579	Glyoxylate/succinic semialdehyde reductase 1
1.10.1300.10	1262	Phosphodiesterase 8, isoform A
1.10.1540.10	596	CDNA FLJ16600 fis, clone TESTI4006704
1.10.220.10	3807	Annexin A5 (Anchorin CII)
1.10.287.130	38277	Methylmalonic aciduria type A protein, mitochondrial
1.10.40.40	45	5'(3')-deoxyribonucleotidase
1.10.540.10	15764	Medium-chain-specific acyl-CoA dehydrogenase, mitocondrial

5.4.8 Analysis of variants at selected domain interfaces

The analysis of variants mapped to interface residues of domain members of three FunFam families from Table 5.6 showing the highest severity weight: Ephrin type-A receptor 2 (ephrin-A2); Matrix metalloproteinase-3 (MMP-3); and T-cell surface glycoprotein CD1e (CD1e); are explored in the next Sub-sections. It is important to notice here that none of the germline and somatic variants identified that cluster into domain interaction interfaces are currently annotated as having disease-association. Since disease-association annotation is likely to lag behind current research in the field, additional queries to the appropriate literature need to be performed to confirm this is the case.

5.4.8.1 Ephrin type-A receptor 2 (ephrin-A2)

Figure 5.12 shows the spatial location of germline and somatic variants mapped to domain 3pix_A_02 (PDB ID 3pix (Kuglstatter et al., 2011), through UniProtKB ID Q06187), which is a member of the Ephrin type-A receptor 2 (ephrin-A2) FunFam family (CATH Superfamily 1.10.510.10, FunFam ID 78758). Domain members of this FunFam include proteins which are involved in cell-signalling and act as receptor kinases. 3pix_A_02 is an epithelial cell receptor tyrosine-kinase domain which was co-crystallised with Bruton's tyrosine kinase (BTK) inhibitor ligand (PDB ligand ID 027) but also participates in a domain-domain interaction with a phosphorylase kinase domain (Figure 5.12 A).

The overview of the variants that are mapped to this domain is shown in Figure 5.12 B. Several of these variants are mapped to the domain-domain and domain-ligand interface and are highlighted in Figure 5.12 C. Among the variants most likely

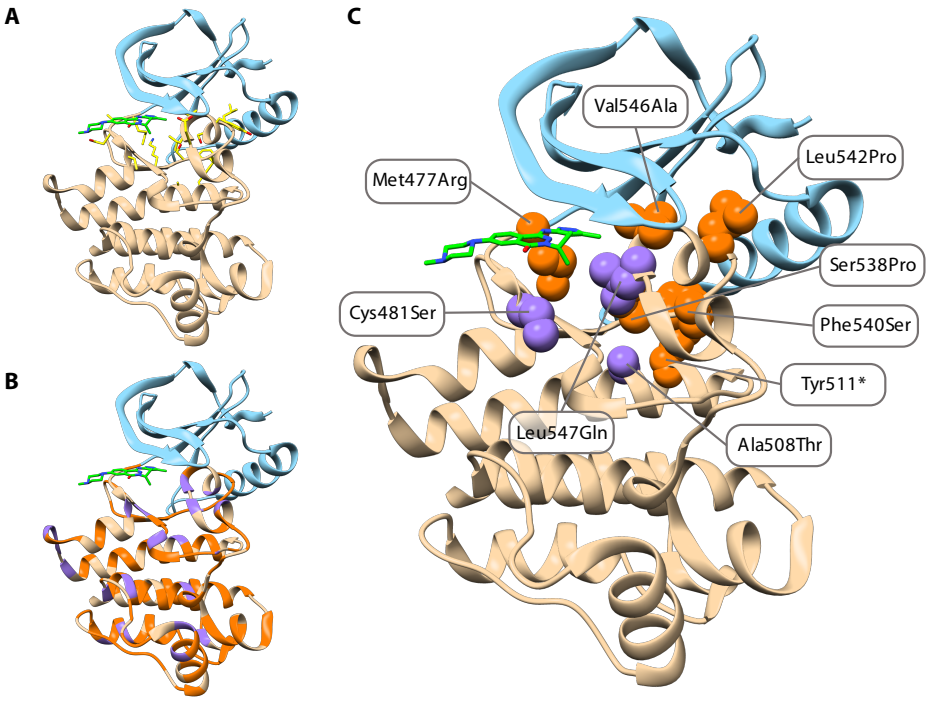


Figure 5.12: Overview of the spatial location of variants mapped to ephrin-A2 interface residues. Variants were mapped to Ephrin type-A receptor 2 (ephrin-A2) (CATH Superfamily 1.10.510.10, FunFam ID 78758), domain 3pix_A_02. Highlight of: A) interface residues and any ligands bound to the domain; B) mapped germline and somatic variants; and C) variants that mapped to structurally conserved interfaces positions. Germline (orange) and somatic (purple) are highlighted. Ligands and the side-chain of interface residues are represented as ball-and-stick, whereas variants that are mapped to interface residues are represented as spheres.

to lead to disruption of the domain-domain interface are Leu542Pro (PhenCode: BTKbase U78027.1:g.66839T>C) and Ser538Pro (PhenCode: BTKbase U78027.1: g.66826T>C), which result in Pro residues, as well as the Tyr511* (PhenCode: BTKbase U78027.1:g.65390C>A) that leads to a truncated protein by encoding a premature stop-codon. Met477Arg (PhenCode: KinMutBase U78027.1:g.65287T>G) might also be important to help to maintain the domain-domain interaction. The analysis of the interaction with the inhibitor ligand indicates that this variation might affect the interaction. Various disease-associated mutations have been identified for domain members of this FunFam. The most severe trait reported for these

mutations is X-linked agammaglobulinemia (XLA, also known as Fleisher syndrome - MIM:307200), which is a humoral immunodeficiency disease that leads to developmental defects in the maturation pathway of B-cells (Conley et al., 1991).

5.4.8.2 Matrix metalloproteinase-3 (MMP-3)

Figure 5.13 shows the spatial location of variants mapped to domain 1hy7_B_00 (PDB ID 1hy7 (Natchus et al., 2001), through UniProtKB ID P08254), which is a member of the Matrix metalloproteinase-3 (MMP-3) FunFam family (CATH Superfamily 3.40.390.10, FunFam ID 13247). This FunFam family is comprised of many metalloproteinase proteins which have proteinase activity (EC. 3.4.24.17) acting as a collagen-activation protein. Figure 5.13 A highlights the various residues involved in the interaction with Ca^{2+} and Zn^{2+} ions, as well as the MMP inhibitor (PDB ligand ID MBS). Figure 5.13 B overviews the spatial location of germline and somatic variants mapped to the domain. Figure 5.13 C highlights those variants that affect domain-ligand interface residues. Among the germline and somatic variants that most likely affect both the coordination of the ions are Gly661Arg (dbSNP: rs782745338), which corresponds to the mutation of a Gly to Arg. This is likely to affect binding to Ca^{2+} as Gly is substituted by a much larger basic amino acid. Gly673Val (COSMIC: COSM1492416) also results in a mutation from Gly, but since Val is a small hydrophobic amino acid, the exchange might not affect the binding to Ca^{2+}. Glu684Lys (COSMIC: COSM3868291) corresponds to a somatic mutation from an acidic amino acid to basic and is likely to affect binding to Ca^{2+}. His666fs* (dbSNP: rs782137879) leads to a frameshift mutation, which is likely to disrupt the

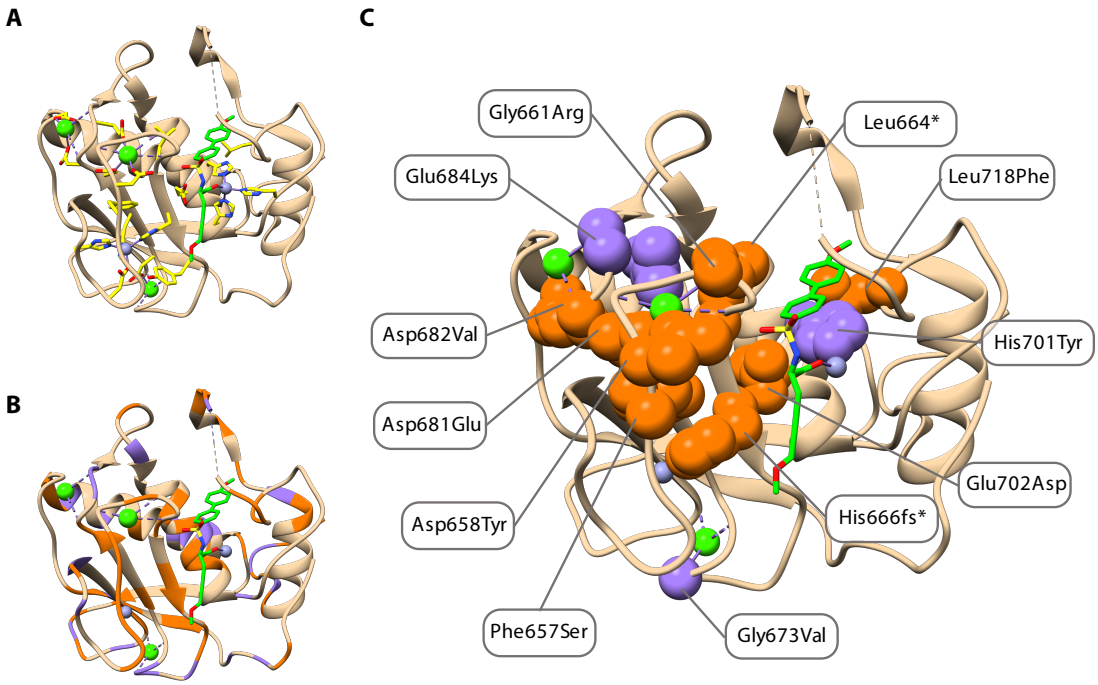


Figure 5.13: Analysis of germline and somatic variants mapped to MMP-3 interface residues. Variants were mapped to Matrix metalloproteinase-3 (MMP-3) (CATH Superfamily 3.40.390.10, FunFam ID 13247), domain 1hy7_B_00. Highlight of: A) interface residues and any ligands bound to the domain; B) mapped germline and somatic variants; and C) variants that mapped to structurally conserved interfaces positions. Germline (orange) and somatic (purple) are highlighted. Ligands and the side-chain of interface residues are represented as ball-and-stick, whereas variants that are mapped to interface residues are represented as spheres.

stability of the entire protein leading to total or partial impairment of its function. Likewise, Leu664* (dbSNP: rs782500546) results in a premature stop and the truncation of the domain.

5.4.8.3 T-cell surface glycoprotein CD1e (CD1e)

Figure 5.14 shows the spatial location of germline and somatic variants mapped to domain 1onq_A_01 (PDB ID 1onq (Zajonc et al., 2003), through UniProtKB ID P06126). This domain is a member of the T-cell surface glycoprotein CD1e (CD1e) FunFam family (CATH Superfamily 3.30.500.10, FunFam ID 3472), which presents

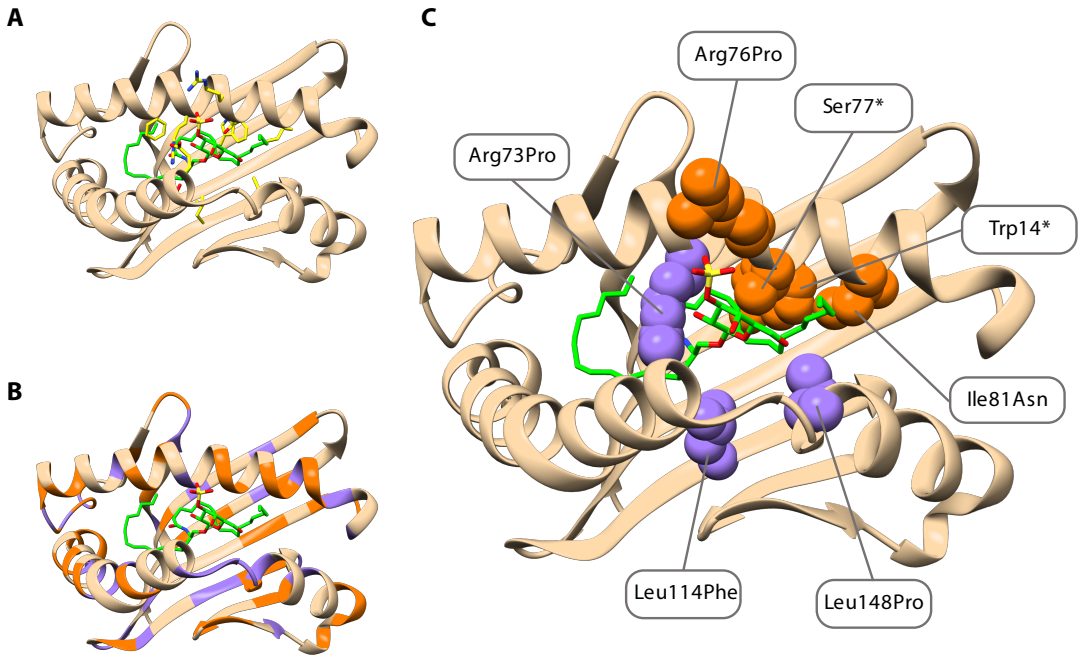


Figure 5.14: Overview of the spatial location of variants mapped to CD1e interface residues. Variants were mapped to T-cell surface glycoprotein CD1e (CD1e) (CATH Superfamily 3.30.500.10, FunFam ID 3472), domain 1onq_A_01. Highlight of: A) interface residues and any ligands bound to the domain; B) mapped germline and somatic variants; and C) variants that mapped to structurally conserved interfaces positions. Germline (orange) and somatic (purple) are highlighted. Ligands and the side-chain of interface residues are represented as ball-and-stick, whereas variants that are mapped to interface residues are represented as spheres.

a Murine Class I Major Histocompatibility Complex, H2-DB, subunit A, domain. Domain members of this FunFam include membrane proteins CD1 antigens involved in the presentation of peptide antigens to the immune system. These bind a variety of self and foreign lipid and glycolipid antigens for presentation to CD1-restricted T cell receptors (TCRs). Figure 5.14 A highlights the CD1e binding pocket with bond sulfatide self-antigen (PDB ID SLF), whereas Figure 5.14 B overviews the variants that are mapped to this domain. Figure 5.14 C highlights those variants that potentially affect domain-ligand binding. Among the variants that likely affect the domain as a whole are Ser77* (dbSNP: rs538916791) and Trp14* (dbSNP: rs149019370) that

lead to a premature truncation of the protein. A total of three variants lead to mutations to Pro, and likely affect binding of the sulfatide self-antigen. These include the germline variant Arg76Pro (dbSNP: rs146048507), and two somatic variants Arg73Pro (COSMIC: COSM1498715) and Leu148Pro (COSMIC: COSM1661050).

5.5 Conclusions

In depth characterisation of variants across different environments by type of substitutions and annotation is important as it enables finding transitions that might be implicated in disease, and that can potentially affect protein stability, activity and function. Therefore, this Chapter focused on the overall analysis of genetic variants at structurally conserved interface positions. The overall analysis of genetic variants at structurally conserved interface positions follows the same general trends observed when investigating genetic variants across all domain sites. The trends of variation were used to help prioritise variants and protein families for further analysis.

The main conclusions of the work presented in this Chapter are:

- Domain-domain interaction types were analysed in the context of CATH SC and FunFams.
- Domain-domain interactions were classified by orientation as a measure of iRMSD clustering.
- A high number of genetic variants were mapped to protein interfaces for FunFam protein families.

- 73-78% of the variants in the three classes mapped to domain-domain interfaces, whereas 13-18% of the variants mapped to domain-ligand interfaces.
- Domain-domain contact propensity matrices were generated and showed that favourable interactions are performed by charged (ionic), hydrophobic and aliphatic residues.
- The frequency of amino acids in the domain-domain interaction dataset follows the same overall trend observed for the abundance of each amino acid in the human proteome.
- Hydrophobic interactions account for 68% of all interactions, followed by hydrogen-bond (15%) and salt-bridge (7%) interactions.
- Mutations to Pro and Arg are frequently observed for disease-associated variants. This is accompanied by mutation of Gly and Cys residues.
- Overall, small differences are observed when comparing both the variation at interfaces and variation at all structural environments, for germline and somatic variants.
- Mutations from charged residues are more frequently observed for somatic variants, whereas mutations from Ile and Val are enriched for somatic variants.
- Regarding the physicochemical properties of the amino acids, disruptive transitions are more consistently observed for disease-associated variants.
- The bigger frequency differences observed between germline and somatic variants are observed for exchanges from sulphur-containing residues to chemically

complex aliphatic and aromatic residues.

- The most problematic amino acid exchanges are those that: disrupt intermolecular bonds; lead to substantial size changes; lead to changes in ionisation state; lead to changes in hydrophobicity; or mutate from/to Pro and Gly.

- Preliminary analysis of some proof-of-concept examples indicates that many variants currently annotated and thought to be neutral might affect protein binding and in that way disrupt important functional interactions.

Chapter 6

Conclusions and future work

6.1 Summary

The work presented in this Thesis focused on the analysis of genetic variation in protein domain families and interaction interfaces. This has been approached by: 1) developing integrative methods for the analysis of structural protein-protein and protein-ligand interactions; 2) combining structural and sequence data to enable the study of genetic variants on protein interfaces; 3) improving the current methods for comprehensive analysis of protein structural and functional families; 4) increasing the availability and understanding of genetic variation at domain interaction interfaces. This chapter summarises the contributions made in each Chapter and provides suggestions of possible future work.

6.2 Development and contents of ProIntVar

Chapter 2 described the development of ProIntVar and overviewed the tools and methods necessary for the analysis of genetic variants in the context of feature-rich protein structural data. Chapter 2 also overviewed the datasets collected and organised in ProIntVar. The main features of ProIntVar are that: 1) it implements routines for generating biological assemblies, defining protein interactions, annotating additional structural features in sites and regions; 2) it allows protein structure, protein sequence and genomic DNA sequence to be cross-mapped; 3) it incorporates generated structure-based MSAs for CATH structural clusters and functional families, which were further extended with similar protein sequences; 4) it allows the comprehensive analysis of protein domain families and exploring protein interfaces across different protein families within a structural perspective; and finally 5) it provides an integrative environment for the analysis of large genetic variation datasets in the context of proteins structure and protein domain families.

The availability of structural data and genetic variation is likely to increase at a steady pace over the next years. Additionally, one of the biggest problems currently in bioinformatics is that many tools stop being updated very shortly after development and release. With both of these points in mind, ProIntVar has been developed in a way that enables updates to the raw structure and sequence datasets which ProIntVar is built on. Several scripts have also been developed to enable the update of such datasets and to perform all the computations needed for populating the database. These cover all the steps necessary from pre-processing the data to generating the analysis results provided in the ProIntVar web-server, which will be

made available at a later stage. This process could in principle be further improved so that an updating 'job' could be scheduled to run on a regular basis. Further improvements to the web-interface could also be made in order to give a wider community of biologists and bioinformatics tools developers access to the computational results. For this, a RESTful API could be developed in a way that flexible web-server endpoints would return a variety of results. Some of the most basic and frequently queried results are already provided in such an API, but the improvement and extensibility of the current methods would be very useful.

6.3 Structural alignment of domain families

In order to perform an enriched analysis of genetic variants under a structural and evolutionary perspective (Chapter 4 and 5), structure-based MSAs were generated for CATH structural and functional families by STAMP. Chapter 3 focused on improving the quality of the generated MSAs by: 1) developing methods that take advantage of the features of STAMP; and 2) exploring the power of HMMs for extending the alignments and annotating them with homologous protein sequences. Both these aspects are key for increasing the scope and reliability of the structure/sequence data currently available for analysis. A new approach for selecting seed domains from a pool of all-against-all domain STAMP scanning superimpositions was developed. This optimisation of the set of transformations obtained for the seed domain led to an overall improvement of the Sc and RMSD alignment measures for 46% of the CATH domain families. Although STAMP produces overall reliable MSAs for CATH SCs and FunFams, the quality of the FunFam-based alignments is

generally higher, which results from a higher structural similarity between the family domain members. Regarding the reliability of the structural alignments, alternative quality assessment measures could be used (Edgar, 2010; Raghava et al., 2003). This would improve the scope of the variation analysis by increasing the confidence with which key functional sites can be inferred across the MSAs. Additionally, it would be useful to investigate the reliability and quality of MSAs generated by new multiple structure alignment programs that have been recently developed, among which: MICAN (Minami et al., 2013), POSA (Li et al., 2014), and UniAlign (Zhao and Sacan, 2015).

Another key related aspect is the aim of increasing the coverage of the structural data available, in order to be able to map/infer potential interaction or variant sites in homologous sequences. It has been predicted that only about half of the human genome is structurally covered (Xie and Bourns, 2005; Marsden et al., 2007; Khafizov et al., 2014). Additionally, both the number of known folds and types of domain-domain interactions that have been identified is believed to be far from complete (Garma et al., 2012). Trying to address this issue, HMM-based approaches have been developed so that CATH domain predictions are provided for many proteomes (Lees et al., 2012). Gene3D (Lees et al., 2012) provides predicted domains that could be mapped to available protein structures not currently covered in CATH. These could be used to extend the STAMP structure-based MSAs or as an alternative to the HMM-based method implemented in this work (Section 3.4.3). Importantly, HMM profiles built from FunFams seed alignments were also recently made available for function prediction and protein annotation in a new protocol called FunFHMMer (Das et al., 2015b). FunFHMMer provides functional family (FunFam)-based

domain assignments and constitutes an alternative to Gene3D. Both structure and (predicted) sequence domains are now included in FunFHMMer, which enables the analysis of domains and particularly domain interactions in the context of a larger protein universe. This helps in increasing the CATH coverage of the current set of protein structures in the PDB, as well as in the prediction/classification of domains in newly solved structures or entire proteomes. As a result of these developments, FunFHMMer could be used in addition to Gene3D, and as an alternative way to extend the STAMP generated MSAs.

The structural data has powered new advances in computational prediction of protein-protein interactions on a genome-wide scale (Zhang et al., 2012; Tuncbag et al., 2011). Owing to the explosive growth in available sequences, novel methods were also developed to predict the 3D structure of proteins (Marks et al., 2012; Hopf et al., 2012, 2014). Experimental methods are also in active continuous development towards better determination of the structure of proteins (Shi et al., 2013; Nannenga and Gonen, 2014) and better identification of native protein complexes (Ewing et al., 2007; Gingras et al., 2007; Kirkwood et al., 2013). The current combination of detailed domain classification and sophisticated HMM domain assignment provides a good source of homology relevant information. Nevertheless, alternative approaches that explore structural homology by using comparative protein structure models exist and their use could be further investigated. Such methods include ModBase (Pieper et al., 2011), IBIS (Inferred Biomolecular Interaction Server) (Shoemaker et al., 2012), Phyre2 (Kelley et al., 2015), and SwissModel (Biasini et al., 2014). New approaches that combine these developments could be valuable for increasing the structural coverage of the current variation analysis.

6.4 Variation across protein domain families

Chapter 4 focused on the overall analysis of genetic variants in protein families, in the context of various structure regions. The in-depth characterisation of variants across different environments is key for finding transitions that might be implicated in disease, and that can potentially affect protein stability, activity and function. The global trends identified are also important to select priority variants and protein families for further analysis. The reason for classifying structure domains into Families and Superfamilies is to be able to compare similar domains within a specific Superfamily/Family to discover the properties of each group. This leads to an inherently high level of redundancy, which which is kept by working with all the domain members of the CATH SCs and FunFams. A domain belonging to a particular Superfamily/Family could thus in principle be chosen to represent the group if the majority of the domains within the group are essentially the same. Redundancy removal approaches for the analysis of nsSNPs in protein structures have been performed by others (David et al., 2012; Yates and Sternberg, 2013b; David and Sternberg, 2015). This process is useful for removing data duplication, for removing duplicated data, but presents a new problem in the form of loss of depth in the analysis of genetic variation within species. For example, in David et al., 2012, the analysis of SNPs using non-redundant structure datasets led to a total of 4,532 nsSNPs in 537 protein structures.

A good future development for ProIntVar would be the analysis of variation at key sites in proteins, such as protein post-translational modified (PTM) sites

(Craveur et al., 2014; Gray et al., 2014; Beltrao et al., 2012). PTMs such as phosphorylation, glycosylation, methylation, acetylation, lipid modifications and ubiquitylation, have wide-ranging effects on protein function and interactions with other molecules and are thereby central to cellular behaviour and responses. PTMs have no effect on protein fold, but have a role in protein function or localisation within the cell, on the cell membrane or in the extracellular matrix. A missense mutation may abolish a PTM site by introducing an amino acid that cannot be modified, or altering the neighbouring residues so that the PTM site cannot be recognised, thereby leading to abnormal protein function. A missense mutation-induced gain of PTM is another possible disease mechanism, leading to protein destabilisation, changes in protein interactions, catalytic properties or other protein functions.

With the amount of genetic variation data growing exponentially, several groups have started performing comparative functional analysis of the distribution of variants between and within protein families and started analysing the specific load of variants at various frequencies in particular biological pathways. Accordingly, it has been found recently that particular members of 2R-ohnologue protein families show some mutation-load skew to a particular member, which is hypothesised to be particularly relevant in cancer (Tinti et al., 2014). Preferential distribution skew has also been identified for mutations in certain protein families of the human kinome (Izarzugaza et al., 2011). This approach would enable finding protein families that are significantly over-mutated or mutation-free (Yates and Sternberg, 2013a; Petrovski et al., 2015). Following these recent developments in the functional characterisation of genetic variation data, it would be interesting to look for both

hot-spots and cold-spots for genetic variation in protein families or specific members of such families. This would enable finding protein families that are significantly over-mutated (polymorphic) or mutation-depleted (protected).

6.5 Variation at conserved interfaces

Chapter 5 focused on the overall analysis of genetic variants at structurally conserved interface positions. Domain-domain interactions were classified by orientation as a measure of iRMSD clustering. The overall analysis of genetic variants at structurally conserved interface positions follows the same general trends observed when investigating genetic variants across all domain sites. The trends of variation were used to help prioritise variants and protein families for further analysis.

Analysis of variation at interaction interfaces could be further explored by measuring the change in binding energetics through the use of modelling and molecular dynamics (Schymkowitz et al., 2005; Kumar and Purohit, 2014; Frappier and Najmanovich, 2014). This would enable cross-checking of transitions that can potentially affect protein stability, activity and function, as a result of either increased or decreased binding affinity. Similarly, the use of newer predictors such as PredicSNP2 (Bendl et al., 2016) could be used to help prioritise the analysis of variants.

Bibliography

Abyzov, A., and Ilyin, V. A. 2007. A comprehensive analysis of non-sequential alignments between all protein structures. *BMC Structural Biology* 7:78.

Acharya, K. R., and Lloyd, M. D. 2005. The advantages and limitations of protein crystal structures. *Trends in Pharmacological Sciences* 26(1):10–14.

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols Human Genetics* Chapter 7:Unit7.20.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* 7(4):248–249.

Ahmad, S., Gromiha, M., Fawareh, H., and Sarai, A. 2004. ASAView: database and tool for solvent accessibility representation in proteins. *BMC Bioinformatics* 5:51.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., et al. 2013. Signatures of mutational processes in human cancer. *Nature* 500(7463):415–421.

Aloy, P., Ceulemans, H., Stark, A., and Russell, R. B. 2003. The Relationship Between Sequence and Interaction Divergence in Proteins. *Journal Molecular Biology* 332(5): 989–998.

Aloy, P., and Russell, R. B. 2006. Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology* 7(3):188–197.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *Journal Molecular Biology* 215(3):403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389–3402.

Alves, R., Chaleil, R. A. G., and Sternberg, M. J. E. 2002. Evolution of enzymes in metabolism: a network perspective. *Journal Molecular Biology* 320(4):751–770.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F. o., Scott, A. F., and Hamosh, A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an Online catalog of human genes and genetic disorders. *Nucleic Acids Research* 43(Database issue):D789–D798.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. 2014. SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Research* 42(Database issue):D310–D314.

Archakov, A. I., Govorun, V. M., Dubanov, A. V., Ivanov, Y. D., Veselovsky, A. V., Lewi, P., and Janssen, P. 2003. Protein-protein interactions as a target for drugs in proteomics. *Proteomics* 3(4):380–391.

Argos, P. 1988. An investigation of protein subunit and domain interfaces. *Protein Engineering, Design and Selection* 2(2):101–113.

Arkin, M. R., and Wells, J. A. 2004. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery* 3(4):301–317.

Bahadur, R. P., Chakrabarti, P., Rodier, F., and Janin, J. 2004. A dissection of specific and non-specific protein-protein interfaces. *Journal Molecular Biology* 336(4):943–955.

Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic Acids Research* 28(1):304–305.

Barton, G. J., and Sternberg, M. J. 1987. Evaluation and improvements in the automatic alignment of protein sequences. *Protein Engineering, Design and Selection* 1(2):89–94.

Barton, G. J. University of Dundee, UK. 2004. OC - A Cluster Analysis Program. `http://www.compbio.dundee.ac.uk/downloads/oc/`. Accessed: 2014-01-18.

Bashton, M., and Chothia, C. 2002. The geometry of domain combination in proteins. *Journal Molecular Biology* 315(4):927–939.

Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O., and Gursoy, A. 2014. PRISM: A web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. *Nucleic Acids Research* 42(Web Server issue):W285–W289.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. 2004. The Pfam protein families database. *Nucleic Acids Research* 32(Database issue):D138–D141.

Bayat, A. 2002. Science, medicine, and the future: Bioinformatics. *British Medical Journal* 324(7344):1018–1022.

de Beer, T. A. P., Laskowski, R. A., Parks, S. L., Sipos, B., Goldman, N., and Thornton, J. M. 2013. Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset. *PLoS Computational Biology* 9(12):e1003382.

Beltrao, P., Albanèse, V., Kenner, L. R., Swaney, D. L., Burlingame, A., Villén, J., Lim, W. A., Fraser, J. S., Frydman, J., and Krogan, N. J. 2012. Systematic functional prioritization of protein posttranslational modifications. *Cell* 150(2):413–425.

Bendl, J., Musil, M., Stourac, J., Zendulka, J., Damborsky, J., and Brezovsky, J. 2016. PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. *PLoS Computational Biology* 12(5):e1004962.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J., and Damborsky, J. 2014. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Computational Biology* 10(1):1–11.

Berg, C., Hedrum, A., Holmberg, A., Pontén, F., Uhlen, M., and Lundeberg, J. 1995. Direct solid-phase sequence analysis of the human p53 gene by use of multiplex polymerase chain reaction and alpha-thiotriphosphate nucleotides. *Clinical Chemistry* 41(10):1461–1466.

Berggård, T., Linse, S., and James, P. 2007a. Methods for the detection and analysis of protein-protein interactions. *Proteomics* 7(16):2833–2842.

Berggård, T., Linse, S., and James, P. 2007b. Methods for the detection and analysis of protein-protein interactions. *Proteomics* 7(16):2833–2842.

Berliner, N., Teyra, J., Çolak, R., Lopez, S. G., and Kim, P. M. 2014. Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS ONE* 9(9):e107353.

Berman, H., Henrick, K., and Nakamura, H. 2003. Announcing the worldwide Protein Data Bank. *Nature Structural Biology* 10(12):980–980.

Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research* 35(Database issue):D301–D303.

Bernal, A., Ear, U., and Kyrpides, N. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research* 29(1):126–127.

Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J., and Poupon, A. 2008. DiMoVo: A Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics* 24(5):652–658.

Bhaskara, R. M., and Srinivasan, N. 2011. Stability of domain structures in multi-domain proteins. *Scientific Reports* 1:40.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., and Schwede, T. 2014. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research* 42(Web Server issue):W252–W258.

Björklund, Å. K., Ekman, D., Light, S., Frey-Skött, J., and Elofsson, A. 2005. Domain rearrangements in protein evolution. *Journal Molecular Biology* 353(4):911–923.

Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E. V., Brunak, S., Jensen, L. J., Nicolae, D., Shah, N. H., Grossman, R. L., Cox, N. J., et al. 2013. A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell* 155(1):70–80.

Blundell, T. L., Sibanda, B. L., Montalvão, R. W., Brewerton, S., Chelliah, V., Worth, C. L., Harmer, N. J., Davies, O., and Burke, D. 2006. Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philosophical transactions of the Royal Society B* 361(1467):413–423.

Bogan, A. A., and Thorn, K. S. 1998. Anatomy of hot spots in protein interfaces. *Journal Molecular Biology* 280(1):1–9.

Bonvin, A. M. 2006. Flexible protein-protein docking. *Current Opinion Structural Biology* 16(2):194–200.

Bordner, A. J., and Abagyan, R. 2005. Statistical analysis and prediction of protein-protein interfaces. *Proteins* 60(3):353–366.

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A., and Weiner, J. 2005. The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences* 62(4):435–445.

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D., and Fitzgerald, P. M. 1997. Macromolecular Crystallographic Information File. *Methods in Enzymology* 277:571–590.

Bradshaw, R. T., Patel, B. H., Tate, E. W., Leatherbarrow, R. J., and Gould, I. R. 2011. Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Protein Engineering Design and Selection* 24(1-2):197–207.

Bresciani, G., Cruz, I. B. M., de Paz, J. A., Cuevas, M. J., and González-Gallego, J. 2013. The MnSOD Ala16Val SNP: relevance to human diseases and interaction with environmental factors. *Free Radical Research* 47(10):781–792.

Briscoe, A. D., Gaur, C., and Kumar, S. 2004. The spectrum of human rhodopsin disease mutations through the lens of interspecific variation. *Gene* 332:107–118.

Brunger, A. T. 1992. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355(6359):472–475.

Buckle, A. M., Cramer, P., and Fersht, A. R. 1996. Structural and energetic responses to cavity-creating mutations in hydrophobic cores: observation of a buried water molecule and the hydrophilic nature of such hydrophobic cavities. *Biochemistry* 35(14):4298–4305.

Burke, D. F., Worth, C. L., Priego, E.-M., Cheng, T., Smink, L. J., Todd, J. A., and Blundell, T. L. 2007. Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics* 8(1):301.

Burley, S. K., and Petsko, G. A. 1985. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* 229(4708):23–28.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S. 2004. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science* 13(1):190–202.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45(10):1113–1120.

Capra, J. A., and Singh, M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics* 23(15):1875–1882.

Capriotti, E., Fariselli, P., and Casadio, R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research* 33(Web Server issue):W306–W310.

Capriotti, E., Nehrt, N. L., Kann, M. G., and Bromberg, Y. 2012. Bioinformatics for personal genome interpretation. *Briefings in Bioinformatics* 13(4):495–512.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* 22(3):231–238.

Carpenter, E. P., Beis, K., Cameron, A. D., and Iwata, S. 2008. Overcoming the challenges of membrane protein crystallography. *Current Opinion Structural Biology* 18(5):581–586.

Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Luigi Martelli, P. 2011. Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Human Mutation* 32(10):1161–1170.

Cavallo, A., and Martin, A. C. R. 2005. Mapping SNPs to protein sequence and structure data. *Bioinformatics* 21(8):1443–1450.

Chasman, D., and Adams, R. M. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *Journal Molecular Biology* 307(2):683–706.

Chen, Y., Cunningham, F., Rios, D., McLaren, W. M., Smith, J., Pritchard, B., Spudich, G. M., Brent, S., Kulesha, E., Marin-Garcia, P., Smedley, D., Birney, E., and Flicek, P. 2010. Ensembl variation resources. *BMC Genomics* 11(1):293.

Cheng, A. Y., Teo, Y. Y., and Ong, R. T.-H. 2014. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics* 30(12):1707–1713.

Cheng, H., Liao, Y., Schaeffer, R. D., and Grishin, N. V. 2015. Manual classification strategies in the ECOD database. *Proteins: Structure, Function and Bioinformatics* 83(April):1238–1251.

Chicurel, M. 2002. Bioinformatics: bringing it all together. *Nature* 419(6908):751–755.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., and Lifton, R. P. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences U.S.A.* 106(45):19096–19101.

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., Harrell, T. M., McMillin, M. J., Wiszniewski, W., Gambin, T., Coban Akdemir, Z. H., Doheny, K., Scott, A. F., Avramopoulos, D., Chakravarti, A., et al. 2015. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics* 97(2):199–215.

Chothia, C., and Janin, J. 1975. Principles of protein-protein recognition. *Nature* 256(5520):705–708.

Chung, J.-L., Wang, W., and Bourne, P. E. 2006. Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins* 62(3):630–640.

Cochrane, G., Karsch-Mizrachi, I., Nakamura, Y., and on behalf of the International Nucleotide Sequence Database Collaboration. 2010. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research* 39(Database):D15–D18.

Codoñer, F. M., and Fares, M. A. 2008. Why Should We Care About Molecular Coevolution? *EMBO Journal* 4:29–38.

Conley, M. E., Burks, A. W., Herrod, H. G., and Puck, J. M. 1991. Molecular analysis of X-linked agammaglobulinemia with growth hormone deficiency. *Journal of Pediatrics* 119(3):392–397.

Craveur, P., Rebehmed, J., and de Brevern, A. G. 2014. PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. *Database* 2014(0):bau041.

Crick, F. 1970. Central dogma of molecular biology. *Nature* 227(5258):561–563.

Cuff, J. A., and Barton, G. J. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Genetics* 40(3):502–511.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Gir n, C. G. a., Gordon, L., Hourlier, T., Hunt, S. E., et al. 2015. Ensembl 2015. *Nucleic Acids Research* 43(Database issue): D662–D669.

Dantzer, J., Moad, C., Heiland, R., and Mooney, S. 2005. MutDB services: Interactive structural analysis of mutation data. *Nucleic Acids Research* 33(Web Server issue): W311–W314.

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G., and Orengo, C. A. 2015a. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics* 31(21):3460–3467.

Das, S., Sillitoe, I., Lee, D., Lees, J. G., Dawson, N. L., Ward, J., and Orengo, C. A. 2015b. CATH FunFHMMer web server: protein functional annotations using functional family assignments. *Nucleic Acids Research* 43(Web Server issue):W148–W153.

David, A., Razali, R., Wass, M. N., and Sternberg, M. J. E. 2012. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human Mutation* 33(2):359–363.

David, A., and Sternberg, M. J. E. 2015. The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease. *Journal Molecular Biology* 427(17):2886–2898.

Davis, F. P., and Sali, A. 2005. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* 21(9):1901–1907.

Dayhoff, M. O., and Schwartz, R. M. 1978. *A model of evolutionary change in proteins, in 'Atlas of Protein Sequence and Structure'*, chap. 22. Elsevier.

Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular and Cellular Proteomics* 1(5):349–356.

Dehouck, Y., Kwasigroch, J. M., Rooman, M., and Gilis, D. 2013. BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Research* 41(Web Server issue):W333–W339.

Dengler, U., Siddiqui, A. S., and Barton, G. J. 2001. Protein structural domains: Analysis of the 3Dee domains database. *Proteins* 42:332–344.

Dobson, R. J., Munroe, P. B., Caulfield, M. J., and Saqi, M. A. 2006. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics* 7:217.

Domagalski, M. J., Zheng, H., Zimmerman, M. D., Dauter, Z., Wlodawer, A., and Minor, W. 2014. The quality and validation of structures from structural genomics. *Methods in Molecular Biology* 1091(Chapter 21):297–314.

Doolittle, R. F. 1981. Similar amino acid sequences: chance or common ancestry? *Science* 214(4517):149–159.

Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics* 14(9):755–763.

Eddy, S. R. 2011. Accelerated profile HMM searches. *PLoS Computational Biology* 7(10): e1002195.

Edgar, R. C. 2010. Quality measures for protein alignment benchmarks. *Nucleic Acids Research* 38(7):2145–2153.

Elokely, K. M., and Doerksen, R. J. 2013. Docking challenge: protein sampling and molecular docking performance. *Journal of Chemical Information and Modeling* 53(8): 1934–1945.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.

Englander, S. W., and Mayne, L. 1992. Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR. *Annual Review of Biophysics and Biomolecular Structure* 21(1):243–265.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., et al. 2007. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology* 3:89.

Exome Aggregation Consortium (ExAC), Cambridge, MA. 2016. Large-scale aggregation of human genomic data. `http://exac.broadinstitute.org`.

Famiglietti, M. L., Estreicher, A., Gos, A., Bolleman, J., Géhant, S., Breuza, L., Bridge, A., Poux, S., Redaschi, N., Bougueleret, L., Xenarios, I., and Consortium, U. 2014. Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Human Mutation* 35(8):927–935.

Fauchère, J. L., Charton, M., Kier, L. B., Verloop, A., and Pliska, V. 1988. Amino acid side chain parameters for correlation studies in biology and pharmacology. *International Journal of Peptide and Protein Research* 32(4):269–278.

Federhen, S. 2012. The NCBI Taxonomy database. *Nucleic Acids Research* 40(Database issue):D136–D143.

Ferrer-Costa, C., Gelpí, J. L., Zamakola, L., Parraga, I., de la Cruz, X., and Orozco, M. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 21(14):3176–3178.

Fields, S. 2005. High-throughput two-hybrid analysis. The promise and the peril. *The FEBS Journal* 272(21):5391–5399.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. 2014a. Pfam: The protein families database. *Nucleic Acids Research* 42(Database issue):D222–D230.

Finn, R. D., Miller, B. L., Clements, J., and Bateman, A. 2014b. IPfam: A database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research* 42(Database issue):D364–D373.

Fischer, E. 1894. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Deutschen Chemischen Gesellschaft* 27(3):2985–2993.

Fiser, A., Simon, I., and Barton, G. J. 1996. Conservation of amino acids in multiple alignments: aspartic acid has unexpected conservation. *FEBS Letters* 397(2-3):225–229.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., et al. 2013. Ensembl 2013. *Nucleic Acids Research* 41(Database issue):D48–D55.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., et al. 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* 43(Database issue):D805–D811.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. 2014. SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* 42(Database issue):D304–D309.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. 2013. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* 41(Database issue):D808–D815.

Frappier, V., and Najmanovich, R. J. 2014. A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. *PLoS Computational Biology* 10(4):e1003569.

Freed-Pastor, W. A., and Prives, C. 2012. Mutant p53: one name, many proteins. *Genes and Development* 26(12):1268–1286.

Gallet, X., Charloteaux, B., Thomas, A., and Brasseur, R. 2000. A fast method to predict protein interaction sites from sequences. *Journal Molecular Biology* 302(4):917–926.

Gao, Y., Wang, R., and Lai, L. 2004. Structure-based method for analyzing protein-protein interfaces. *Journal of Molecular Modeling* 10(1):44–54.

Garma, L., Mukherjee, S., Mitra, P., and Zhang, Y. 2012. How many protein-protein interactions types exist in nature? *PLoS ONE* 7(6):1–9.

Giacomini, K. M., Brett, C. M., Altman, R. B., Benowitz, N. L., Dolan, M. E., Flockhart, D. A., Johnson, J. A., Hayes, D. F., Klein, T., Krauss, R. M., Kroetz, D. L., McLeod, H. L., Nguyen, A. T., Ratain, M. J., Relling, M. V., et al. 2007. The Pharmacogenetics Research Network: From SNP discovery to clinical drug response. *Clinical Pharmacology and Therapeutics* 81(3):328–345.

Giardine, B., Riemer, C., Hefferon, T., Thomas, D., Hsu, F., Zielenski, J., Sang, Y., Elnitski, L., Cutting, G., Trumbower, H., Kern, A., Kuhn, R., Patrinos, G. P., Hughes, J., Higgs, D., et al. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. *Human Mutation* 28(6):554–562.

Gill, G. 2004. SUMO and ubiquitin in the nucleus: different functions, similar mechanisms? *Genes and Development* 18(17):2046–2059.

Gilliland, G. L., and Ladner, J. E. 1996. Crystallization of biological macromolecules for X-ray diffraction studies. *Current Opinion Structural Biology* 6(5):595–603.

Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. 2007. Analysis of protein complexes using mass spectrometry. *Nature Reviews Molecular Cell Biology* 8(8):645–654.

Glaser, F., Steinberg, D. M., Vakser, I. A., and Ben-Tal, N. 2001. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Structure, Function, and Genetics* 43(2):89–102.

Goh, K.-I., and Choi, I.-G. 2012. Exploring the human diseasome: the human disease network. *Briefings in Functional Genomics* 11(6):533–542.

Goldman, N., Thorne, J. L., and Jones, D. T. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149(1):445–458.

Gong, S., and Blundell, T. L. 2010. Structural and functional restraints on the occurrence of single amino acid variations in human proteins. *PLoS ONE* 5(2):e9186.

Gong, S., Yoon, G., Jang, I., Bolser, D., Dafas, P., Schroeder, M., Choi, H., Cho, Y., Han, K., Lee, S., Choi, H., Lappe, M., Holm, L., Kim, S., Oh, D., et al. 2005. PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics* 21(10): 2541–2543.

Gonzaga-Jauregui, C., Lupski, J. R., and Gibbs, R. A. 2012. Human genome sequencing in health and disease. *Annual Review of Medicine* 63(1):35–61.

González-Pérez, A., and López-Bigas, N. 2011. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *The American Journal of Human Genetics* 88(4):440–449.

Gore, S., Velankar, S., and Kleywegt, G. J. 2012. Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallographica Section D* 68(Pt 4):478–483.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal Molecular Biology* 313(4):903–919.

Gray, V. E., Liu, L., Nirankari, R., Hornbeck, P. V., and Kumar, S. 2014. Signatures of natural selection on mutations of residues with multiple posttranslational modifications. *Molecular Biology and Evolution* 31(7):1641–1645.

Gribenko, A. V., Patel, M. M., Liu, J., McCallum, S. A., Wang, C., and Makhatadze, G. I. 2009. Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. *Proceedings of the National Academy of Sciences U.S.A.* 106(8): 2601–2606.

Guharoy, M., and Chakrabarti, P. 2010. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics* 11(1):286.

Gulati, S., Cheng, T. M. K., and Bates, P. A. 2013. Cancer networks and beyond: Interpreting mutations using the human interactome and protein structure. *Seminars in Cancer Biology* 23(4):219–226.

Gutmanas, A., Alhroub, Y., Battle, G. M., Berrisford, J. M., Bochet, E., Conroy, M. J., Dana, J. M., Fernandez Montecelo, M. A., van Ginkel, G., Gore, S. P., Haslam, P., Hatherley, R., Hendrickx, P. M. S., Hirshberg, M., Lagerstedt, I., et al. 2014. PDBe: Protein Data Bank in Europe. *Nucleic Acids Research* 42(Database issue):D285–D291.

Gutteridge, A., and Thornton, J. M. 2005. Understanding nature's catalytic toolkit. *Trends in Biochemical Sciences* 30(11):622–629.

Hadley, C., and Jones, D. T. 1999. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* 7(9):1099–1112.

Hall, S. R. 1991. The STAR file: a new format for electronic data transfer and archiving. *Journal of Chemical Information and Modeling* 31(2):326–333.

Halperin, I., Wolfson, H., and Nussinov, R. 2004. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure* 12(6):1027–1038.

Halperin, I., Wolfson, H., and Nussinov, R. 2006. Correlated mutations: Advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins* 63(4):832–845.

Hamelryck, T. 2009. Probabilistic models and machine learning in structural bioinformatics. *Statistical Methods in Medical Research* 18(5):505–526.

Hamelryck, T., and Manderick, B. 2003. PDB file parser and structure class implemented in Python. *Bioinformatics* 19(17):2308–2310.

Hamosh, A. 2004. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* 33(Database issue):D514–D517.

Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A., and Clarke, J. 2007. The folding and evolution of multidomain proteins. *Nature Reviews Molecular Cell Biology* 8(4): 319–330.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32(Database issue):D258–D261.

Hasegawa, H., and Holm, L. 2009. Advances and pitfalls of protein structural alignment. *Current Opinion Structural Biology* 19(3):341–348.

Hashem, Y., des Georges, A., Dhote, V., Langlois, R., Liao, H. Y., Grassucci, R. A., Pestova, T. V., Hellen, C. U. T., and Frank, J. 2013. Hepatitis-C-virus-like internal ribosome entry sites displace eIF3 to gain access to the 40S subunit. *Nature* 503(7477): 539–543.

Heinzen, E. L., Swoboda, K. J., Hitomi, Y., Gurrieri, F., Nicole, S., de Vries, B., Tiziano, F. D., Fontaine, B., Walley, N. M., Heavin, S., Panagiotakaki, E., European Alternating Hemiplegia of Childhood (AHC) Genetics Consortium, Biobanca e Registro Clinico per l'Emiplegia Alternante (I.B.AHC) Consortium, European Network for Research on Alternating Hemiplegia (ENRAH) for Small and Medium-sized Enterpriese (SMEs) Consortium, Fiori, S., et al. 2012. De novo mutations in ATP1A3 cause alternating hemiplegia of childhood. *Nature Genetics* 44(9):1030–1034.

Henikoff, S., and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences U.S.A.* 89(22):10915–10919.

Henrick, K., Feng, Z., Bluhm, W. F., Dimitropoulos, D., Doreleijers, J. F., Dutta, S., Flippen-Anderson, J. L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C. L., Markley, J. L., Nakamura, H., Newman, R., Shimizu, Y., et al. 2007. Remediation of the protein data bank archive. *Nucleic Acids Research* 36(Database):D426–D433.

Henrick, K., and Thornton, J. M. 1998. PQS: a protein quaternary structure file server. *Trends in Biochemical Sciences* 23(9):358–361.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., et al. 2004. IntAct: an open source molecular interaction database. *Nucleic Acids Research* 32(Database issue):D452–D455.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., and Flicek, P. 2016. Ensembl comparative genomics resources. *Database* 2016:bav096.

Hintze, B. J., Lewis, S. M., Richardson, J. S., and Richardson, D. C. 2016. MolProbity's Ultimate Rotamer-Library Distributions for Model Validation. *Proteins* 84(9):1177–1189.

Holm, L., Kaariainen, S., Rosenstrom, P., and Schenkel, A. 2008. Searching protein structure databases with DaliLite v.3. *Bioinformatics* 24(23):2780–2781.

Holm, L., and Sander, C. 1996. The FSSP database: Fold classification based on structure-structure alignment of proteins. *Nucleic Acids Research* 24(1):206–209.

Holm, L., and Sander, C. 1999. Protein folds and families: Sequence and structure alignments. *Nucleic Acids Research* 27(1):244–247.

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., and Marks, D. S. 2012. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* 149(7):1607–1621.

Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Sander, C., Bonvin, A. M. J. J., and Marks, D. S. 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:65.

Hu, Z., Ma, B., Wolfson, H., and Nussinov, R. 2000. Conservation of polar residues as hot spots at protein interfaces. *Proteins: Structure, Function, and Genetics* 39(4):331–342.

Hubbard, S. J., and Argos, P. 1994. Cavities and packing at protein interfaces. *Protein Science* 3(12):2194–2206.

Hubbard, T. J., Murzin, A. G., Brenner, S. E., and Chothia, C. 1997. SCOP: a Structural Classification Of Proteins database. *Nucleic Acids Research* 25(1):236–239.

Huminiecki, L., and Conant, G. C. 2012. Polyploidy and the evolution of complex traits. *International Journal of Evolutionary Biology* 2012(4):292068–292112.

Hurst, J. M., McMillan, L. E. M., Porter, C. T., Allen, J., Fakorede, A., and Martin, A. C. R. 2009. The SAAPdb web resource: a large-scale structural analysis of mutant proteins. *Human Mutation* 30(4):616–624.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences U.S.A.* 98(8):4569–4574.

Ivanov, A. A., Khuri, F. R., and Fu, H. 2013. Targeting protein–protein interactions as an anticancer strategy. *Trends in Pharmacological Sciences* 34(7):393–400.

Izarzugaza, J. M. G., Hopcroft, L. E. M., Baresic, A., Orengo, C. A., Martin, A. C. R., and Valencia, A. 2011. Characterization of pathogenic germline mutations in human protein kinases. *BMC Bioinformatics* 12 Suppl 4:S1.

Jackson, R. N., McCoy, A. J., Terwilliger, T. C., Read, R. J., and Wiedenheft, B. 2015. X-ray structure determination using low-resolution electron microscopy maps for molecular replacement. *Nature Protocols* 10(9):1275–1284.

Janin, J., and Chothia, C. 1990. The structure of protein-protein recognition sites. *Journal Biological Chemistry* 265(27):16027–16030.

Janin, J. l., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., Vakser, I., and Wodak, S. J. 2003. CAPRI: A critical assessment of PRedicted interactions. *Proteins: Structure, Function, and Genetics* 52(1):2–9.

Jefferson, E. R., Walsh, T. P., and Barton, G. J. 2006. Biological units and their effect upon the properties and prediction of protein-protein interactions. *Journal Molecular Biology* 364(5):1118–1129.

Jefferson, E. R., Walsh, T. P., and Barton, G. J. 2007a. A comparison of SCOP and CATH with respect to domain-domain interactions. *Proteins* 70(1):54–62.

Jefferson, E. R., Walsh, T. P., Roberts, T. J., and Barton, G. J. 2007b. SNAPPI-DB: A database and API of structures, iNterfaces and Alignments for Protein-Protein Interactions. *Nucleic Acids Research* 35(Database issue):D580–D589.

Johansson, F., and Toh, H. 2010. A comparative study of conservation and variation scores. *BMC Bioinformatics* 11(1):388.

Jones, S., Marin, A., and Thornton, J. M. 2000. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Engineering, Design and Selection* 13(2):77–82.

Jones, S., and Thornton, J. M. 1997. Analysis of protein-protein interaction sites using surface patches. *Journal Molecular Biology* 272(1):121–132.

Jones, S., and Thornton, J. M. 1996. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences U.S.A.* 93(1):13–20.

Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard, M. 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallographica Section A* 47 (Pt 2):110–119.

Jonic, S., and Vénien-Bryan, C. 2009. Protein structure determination by electron cryomicroscopy. *Current Opinion in Pharmacology* 9(5):636–642.

Jonsson, P. F., and Bates, P. A. 2006. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22(18):2291–2297.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577–2637.

Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., Lander, E. S., and Getz, G. 2015. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences U.S.A.* 112(40):E5486–5495.

Kanehisa, M., and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28(1):27–30.

Karchin, R. 2009. Next generation tools for the annotation of human SNPs. *Briefings in Bioinformatics* 10(1):35–52.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 10(6):845–858.

Kenworthy, A. K. 2001. Imaging Protein-Protein Interactions Using Fluorescence Resonance Energy Transfer Microscopy. *Methods* 24(3):289–296.

Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. 2008. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chemical Reviews* 108(4):1225–1244.

Keskin, O., Ma, B., and Nussinov, R. 2005. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal Molecular Biology* 345(5):1281–1294.

Khafizov, K., Madrid-Aliste, C., Almo, S. C., and Fiser, A. 2014. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proceedings of the National Academy of Sciences U.S.A.* 111(10):3733–3738.

Khan, S. H., Ahmad, F., Ahmad, N., Flynn, D. C., and Kumar, R. 2011. Protein-protein interactions: principles, techniques, and their potential role in new drug development. *Journal of Biomolecular Structure and Dynamics* 28(6):929–938.

Khan, S., and Vihinen, M. 2007. Spectrum of disease-causing mutations in protein secondary structures. *BMC Structural Biology* 7(1):56.

Kirkwood, K. J., Ahmad, Y., Larance, M., and Lamond, A. I. 2013. Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. *Molecular and Cellular Proteomics* 12(12):3851–3873.

Knapp, D. R. 1996. Mass Spectrometry in the Biological Sciences A. L. Burlingame and S. A. Carr, Editors . *Journal of the American Society for Mass Spectrometry* 7(7): 692–692.

Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. 2012. Principles for designing ideal protein structures. *Nature* 491(7423): 222–227.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., and Lesk, A. M. 2006. MUSTANG: a multiple structural alignment algorithm. *Proteins* 64(3):559–574.

Kono, H., Yuasa, T., Nishiue, S., and Yura, K. 2007. coliSNP database server mapping nsSNPs on protein structures. *Nucleic Acids Research* 36(Database):D409–D413.

Koshland, D. E. 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences U.S.A.* 44(2):98–104.

Krawczak, M., Ball, E. V., Fenton, I., Stenson, P. D., Abeysinghe, S., Thomas, N., and Cooper, D. N. 2000. Human Gene Mutation Database: a biomedical information and research resource. *Human Mutation* 15(1):45–51.

Krissinel, E., and Henrick, K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D* 60(Pt 12 Pt 1):2256–2268.

Krissinel, E. 2011. Macromolecular complexes in crystals and solutions. *Acta Crystallographica Section D* 67(4):376–385.

Krissinel, E., and Henrick, K. 2005. Multiple alignment of protein structures in three dimensions. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, 67–78. European Bioinformatics Institute, Cambridge, United Kingdom.

Krissinel, E., and Henrick, K. 2007. Inference of macromolecular assemblies from crystalline state. *Journal Molecular Biology* 372(3):774–797.

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324(5924):255–258.

Kuglstatter, A., Wong, A., Tsing, S., Lee, S. W., Lou, Y., Villase or, A. G., Bradshaw, J. M., Shaw, D., Barnett, J. W., and Browner, M. F. 2011. Insights into the conformational flexibility of Bruton's tyrosine kinase from multiple ligand complex structures. *Protein Science* 20(2):428–436.

Kumar, A., and Purohit, R. 2014. Use of Long Term Molecular Dynamics Simulation in Predicting Cancer Associated SNPs. *PLoS Computational Biology* 10(4):e1003318.

Kumar, P., Henikoff, S., and Ng, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* 4(7):1073–1081.

Kumar, S., and Nussinov, R. 2002. Close-range electrostatic interactions in proteins. *ChemBioChem* 3(7):604–617.

Kummerfeld, S. K., and Teichmann, S. A. 2009. Protein domain organisation: adding order. *BMC Bioinformatics* 10(1):39.

Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F., and Mackay, J. P. 2011. Macromolecular NMR spectroscopy for the non-spectroscopist. *The FEBS Journal* 278(5): 687–703.

Kwok, P.-Y., and Duan, S. 2003. SNP discovery by direct DNA sequencing. *Methods in Molecular Biology* 212:71–84.

Lahti, J. L., Tang, G. W., Capriotti, E., Liu, T., and Altman, R. B. 2012. Bioinformatics and variability in drug response: a protein structural perspective. *Journal of The Royal Society Interface* 9(72):1409–1437.

Lander, G. C., Saibil, H. R., and Nogales, E. 2012. Go hybrid: EM, crystallography, and beyond. *Current Opinion Structural Biology* 22(5):627–635.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* 44(Database issue):D862–D868.

Larranaga, P. 2006. Machine learning in bioinformatics. *Briefings in Bioinformatics* 7(1): 86–112.

Launay, G., and Simonson, T. 2008. Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC Bioinformatics* 9(1):427.

Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S. J., Newman, R. H., Oldfield, T. J., Rees, I., Sahni, G., Sala, R., et al. 2011. EMDataBank.org: Unified data resource for CryoEM. *Nucleic Acids Research* 39(Database issue):D456–D464.

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., and Orengo, C. 2012. Gene3D: A domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Research* 40(Database issue): D465–D471.

Lengyel, J., Hnath, E., Storms, M., and Wohlfarth, T. 2014. Towards an integrative structural biology approach: combining Cryo-TEM, X-ray crystallography, and NMR. *Journal of Structural and Functional Genomics* 15(3):117–124.

Li, X., Keskin, O., Ma, B., Nussinov, R., and Liang, J. 2004. Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *Journal Molecular Biology* 344(3):781–795.

Li, Z., Natarajan, P., Ye, Y., Hrabe, T., and Godzik, A. 2014. POSA: A user-driven, interactive multiple protein structure alignment server. *Nucleic Acids Research* 42(Web Server issue):W240–W245.

Lilley, D. M., and Wilson, T. J. 2000. Fluorescence resonance energy transfer as a structural tool for nucleic acids. *Current Opinion in Chemical Biology* 4(5):507–517.

Littler, S. J., and Hubbard, S. J. 2005. Conservation of Orientation and Sequence in Protein Domain-Domain Interactions. *Journal Molecular Biology* 345(5):1265–1279.

Liu, R., Baase, W. A., and Matthews, B. W. 2000. The introduction of strain and its effects on the structure and stability of T4 lysozyme. *Journal Molecular Biology* 295(1): 127–145.

Livingstone, C. D., and Barton, G. J. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Computer Applications in the Biosciences* 9(6):745–756.

Lo Conte, L., Chothia, C., and Janin, J. 1999. The atomic structure of protein-protein recognition sites. *Journal Molecular Biology* 285(5):2177–2198.

Lo Conte, L., Ailey, B., Hubbard, T. J. P., Brenner, S. E., Murzin, A. G., and Chothia, C. 2000. SCOP: A structural classification of proteins database. *Nucleic Acids Research* 28(1):257–259.

Loladze, V. V., Ermolenko, D. N., and Makhatadze, G. I. 2002. Thermodynamic Consequences of Burial of Polar and Non-polar Amino Acid Residues in the Protein Interior. *Journal Molecular Biology* 320(2):343–357.

Luca, S., Heise, H., and Baldus, M. 2004. High-Resolution Solid-State NMR Applied to Polypeptides and Membrane Proteins. *ChemInform* 35(6).

Lupyan, D., Leo-Macias, A., and Ortiz, A. R. 2005. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* 21(15):3255–3263.

Luu, T.-D., Rusu, A.-M., Walter, V., Ripp, R., Moulinier, L., Muller, J., Toursel, T., Thompson, J. D., Poch, O., and Nguyen, H. 2012. MSV3d: database of human MisSense Variants mapped to 3D protein structure. *Database* 2012:bas018.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. 2003. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences U.S.A.* 100(10):5772–5777.

Ma, Q., and Lu, A. Y. H. 2011. Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacological Reviews* 63(2):437–459.

Madej, T., Lanczycki, C. J., Zhang, D., Thiessen, P. A., Geer, R. C., Marchler-Bauer, A., and Bryant, S. H. 2014. MMDB and VAST+: Tracking structural similarities between macromolecular complexes. *Nucleic Acids Research* 42(Database issue):D297–D303.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. 2011. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research* 39(Database issue):D52–D57.

Maleki, M., Hall, M., and Rueda, L. 2013. Using desolvation energies of structural domains to predict stability of protein complexes. *Network modeling analysis in health informatics and bioinformatics* 2(4):267–275.

Mao, B., Tejero, R., Baker, D., and Montelione, G. T. 2014. Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *Journal of the American Chemical Society* 136(5):1893–1906.

Marks, D. S., Hopf, T. A., and Sander, C. 2012. Protein structure prediction from sequence variation. *Nature Biotechnology* 30(11):1072–1080.

Marsden, R. L., Lewis, T. A., and Orengo, C. A. 2007. Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics* 8(1):86.

Marsh, J. A., and Teichmann, S. A. 2015. Structure, dynamics, assembly, and evolution of protein complexes. *Annual Review of Biochemistry* 84(1):551–575.

Martin, A. C. R., Facchiano, A. M., Cuff, A. L., Hernandez-Boussard, T., Olivier, M., Hainaut, P., and Thornton, J. M. 2002. Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. *Human Mutation* 19(2):149–164.

Maskos, U., and Southern, E. M. 1992. Oligonucleotide hybridisations on glass supports: a novel linker for oligonucleotide synthesis and hybridisation properties of oligonucleotides synthesised in situ. *Nucleic Acids Research* 20(7):1679–1684.

Matthews, B. W. 1993. Structural and genetic analysis of protein stability. *Annual Review of Biochemistry* 62(1):139–160.

McCoy, A. J., Chandana Epa, V., and Colman, P. M. 1997. Electrostatic complementarity at protein/protein interfaces. *Journal Molecular Biology* 268(2):570–584.

McDonald, I. K., and Thornton, J. M. 1994. Satisfying hydrogen bonding potential in proteins. *Journal Molecular Biology* 238(5):777–793.

Menke, M., Berger, B., and Cowen, L. 2008. Matt: Local flexibility aids protein multiple structure alignment. *PLoS Computational Biology* 4(1):88–99.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417(6887):399–403.

Miller, M. P., and Kumar, S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Human Molecular Genetics* 10(21):2319–2328.

Minami, S., Sawada, K., and Chikenji, G. 2013. MICAN: a protein structure alignment algorithm that can handle Multiple-chains, Inverse alignments, C(α) only models, Alternative alignments, and Non-sequential alignments. *BMC Bioinformatics* 14(1):24.

Mintseris, J., and Weng, Z. 2003. Atomic Contact Vectors in Protein-Protein Recognition. *Proteins: Structure, Function, and Genetics* 53(3):629–639.

Mirnezami, R., Nicholson, J., and Darzi, A. 2012. Preparing for Precision Medicine. *New England Journal of Medicine* 366(6):489–491.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.-Y., Bateman, A., et al. 2015. The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Research* 43(Database issue): D213–D221.

Mooney, S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics* 6(1):44–56.

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. 2014. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research* 42(Database issue):D374–D379.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

Mullard, A. 2012. Protein-protein interaction inhibitors get into the groove. *Nature Reviews Drug Discovery* 11(3):173–175.

Murzin, A. G., and Bateman, A. 1997. Distant homology recognition using structural classification of proteins. *Proteins: Structure, Function, and Genetics* 29(Database issue):D105–D112.

Nannenga, B. L., and Gonen, T. 2014. Protein structure determination by MicroED. *Current Opinion Structural Biology* 27:24–31.

Natchus, M. G., Bookland, R. G., Laufersweiler, M. J., Pikul, S., Almstead, N. G., De, B., Janusz, M. J., Hsieh, L. C., Gu, F., Pokross, M. E., Patel, V. S., Garver, S. M., Peng, S. X., Branch, T. M., King, S. L., et al. 2001. Development of new carboxylic acid-based MMP inhibitors derived from functionalized propargylglycines. *Journal of Medicinal Chemistry* 44(7):1060–1071.

Nei, M., Suzuki, Y., and Nozawa, M. 2010. The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics* 11(1):265–289.

Nelson, M. R., Marnellos, G., Kammerer, S., Hoyal, C. R., Shi, M. M., Cantor, C. R., and Braun, A. 2004. Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Research* 14(8):1664–1668.

Ng, P. C., and Henikoff, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research* 31(13):3812–3814.

Ng, P. C., and Henikoff, S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics* 7(1):61–80.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. 2009. Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics* 42(1):30–35.

NHLBI GO Exome Sequencing Project (ESP), Seattle, WA. 2016. Exome Variant Server. `http://evs.gs.washington.edu/EVS/`.

Nickell, S., Kofler, C., Leis, A. P., and Baumeister, W. 2006. A visual approach to proteomics. *Nature Reviews Molecular Cell Biology* 7(3):225–230.

Niedzialkowska, E., Gasiorowska, O., Handing, K. B., Majorek, K. A., Porebski, P. J., Shabalin, I. G., Zasadzinska, E., Cymborowski, M., and Minor, W. 2016. Protein purification and crystallization artifacts: The tale usually not told. *Protein Science* 25(3):720–733.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* 12(6):443–451.

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., Wuchty, S., and Panchenko, A. R. 2013. Cancer Missense Mutations Alter Binding Properties of Proteins and Their Interaction Networks. *PLoS ONE* 8(6):e66273.

Nooren, I. M., and Thornton, J. M. 2003a. Structural characterisation and functional significance of transient protein-protein interactions. *Journal Molecular Biology* 325(5): 991–1018.

Nooren, I. M. A., and Thornton, J. M. 2003b. Diversity of protein-protein interactions. *EMBO Journal* 22(14):3486–3492.

Prado-Montes de Oca, E., Velarde-Félix, J. S., Ríos-Tostado, J. J., Picos-Cárdenas, V. J., and Figuera, L. E. 2009. SNP 668C (-44) alters a NF-kappaB1 putative binding site in non-coding strand of human beta-defensin 1 (DEFB1) and is associated with lepromatous leprosy. *Infection, Genetics and Evolution* 9(4):617–625.

Ofran, Y., and Rost, B. 2003. Analysing six types of protein-protein interfaces. *Journal Molecular Biology* 325(2):377–387.

Okada, K., Kanaya, S., and Asai, K. 2005. Accurate extraction of functional associations between proteins based on common interaction partners and common domains. *Bioinformatics* 21(9):2043–2048.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., et al. 2014. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* 42(Database issue):D358–D363.

Orengo, C. A., Jones, D. T., and Thornton, J. M. 1994. Protein superfamilies and domain superfolds. *Nature* 372(6507):631–634.

Orengo, C. A., and Taylor, W. R. 1996. SSAP: sequential structure alignment program for protein structure comparison. *Methods in Enzymology* 266:617–635.

Orengo, C. A., Bray, J. E., Buchan, D. W. A., Harrison, A., Lee, D., Pearl, F. M. G., Sillitoe, I., Todd, A. E., and Thornton, J. M. 2002. The CATH protein family database: A resource for structural and functional annotation of genomes. *Proteomics* 2(1):11–21.

Orengo, C. A., and Thornton, J. M. 2005. Protein families and their evolution-a structural perspective. *Annual Review of Biochemistry* 74(1):867–900.

Ouzounis, C. A., and Valencia, A. 2003. Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics* 19(17):2176–2190.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., Markowitz, V. M., and Kyrpides, N. C. 2012. The Genomes OnLine Database (GOLD) v.4: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* 40(Database issue):D571–D579.

Pajunen, M., Turakainen, H., Poussu, E., Peränen, J., Vihinen, M., and Savilahti, H. 2007. High-precision mapping of protein protein interfaces: an integrated genetic strategy combining en masse mutagenesis and DNA-level parallel analysis on a yeast two-hybrid platform. *Nucleic Acids Research* 35(16):e103.

Panchenko, A. R., Kondrashov, F., and Bryant, S. 2004. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Science* 13(4):884–892.

Panigrahi, S. K., and Desiraju, G. R. 2007. Strong and weak hydrogen bonds in the protein-ligand interface. *Proteins* 67(1):128–141.

Park, J., Lappe, M., and Teichmann, S. A. 2001. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *Journal Molecular Biology* 307(3):929–938.

Patwardhan, A., Ashton, A., Brandt, R., Butcher, S., Carzaniga, R., Chiu, W., Collinson, L., Doux, P., Duke, E., Ellisman, M. H., Franken, E., Grünewald, K., Heriche, J.-K., Koster, A., Kühlbrandt, W., et al. 2014. A 3D cellular context for the macromolecular world. *Nature Structure Molecular Biology* 21(10):841–845.

Pazos, F., and Valencia, A. 2008. Protein co-evolution, co-adaptation and interactions. *EMBO Journal* 27(20):2648–2655.

Peterson, T. A., Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A., and Kann, M. G. 2010. DMDM: domain mapping of disease mutations. *Bioinformatics* 26(19):2458–2459.

Petrovski, S., Gussow, A. B., Wang, Q., Halvorsen, M., Han, Y., Weir, W. H., Allen, A. S., and Goldstein, D. B. 2015. The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLoS Genetics* 11(9):e1005492.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. 2004. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25(13):1605–1612.

Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E. C., Pettersen, E. F., Huang, C. C., Datta, R. S., Sampathkumar, P., Madhusudhan, M. S., Sjölander, K., Ferrin, T. E., et al. 2011. ModBase,a database of annotated comparative protein structure models,and associated resources. *Nucleic Acids Research* 39(Database issue):D465–D474.

Ponting, C. P., and Russell, R. R. 2002. The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure* 31:45–71.

Porta-Pardo, E., Hrabe, T., and Godzik, A. 2015. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Research* 43(Database issue):D968–D973.

Porter, C. T., Bartlett, G. J., and Thornton, J. M. 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research* 32(Database issue):D129–D133.

Poussu, E. 2005. A gene truncation strategy generating N- and C-terminal deletion variants of proteins for functional studies: mapping of the Sec1p binding domain in yeast Mso1p by a Mu in vitro transposition-based approach. *Nucleic Acids Research* 33(12):e104.

Powers, R., Copeland, J. C., Germer, K., Mercier, K. A., Ramanathan, V., and Revesz, P. 2006. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* 65(1):124–135.

Prives, C., and Hall, P. A. 1999. The p53 pathway. *The Journal of Pathology* 187(1): 112–126.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. 2003. NCBI Reference Sequence project: Update and current status. *Nucleic Acids Research* 31(1):34–37.

Putnam, C. D., Hammel, M., Hura, G. L., and Tainer, J. A. 2007. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Reviews of Biophysics* 40(3):191–285.

R Core Development Team. 2016. R: A language and environment for statistical computing. Accessed: 2025-01-01.

Raghava, G. P. S., and Barton, G. J. 2006. Quantification of the variation in percentage identity for protein sequence alignments. *BMC Bioinformatics* 7:415.

Raghava, G. P. S., Searle, S. M. J., Audley, P. C., Barber, J. D., and Barton, G. J. 2003. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics* 4(1):47.

Ramensky, V., Bork, P., and Sunyaev, S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research* 30(17):3894–3900.

Redfern, O. C., Harrison, A., Dallman, T., Pearl, F. M. G., and Orengo, C. A. 2007. CATHEDRAL: A fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Computational Biology* 3(11):2333–2347.

Reeves, G. A., Dallman, T. J., Redfern, O. C., Akpor, A., and Orengo, C. A. 2006. Structural diversity of domain superfamilies in the CATH database. *Journal Molecular Biology* 360(3):725–741.

Rios, D., McLaren, W. M., Chen, Y., Birney, E., Stabenau, A., Flicek, P., and Cunningham, F. 2010. A database and API for variation, dense genotyping and resequencing data. *BMC Bioinformatics* 11(1):238.

Rivoire, O., and Leibler, S. 2014. A model for the generation and transmission of variations in evolution. *Proceedings of the National Academy of Sciences U.S.A.* 111(19):E1940–1949.

Roos, D. S. 2001. Bioinformatics-Trying to Swim in a Sea of Data. *Science* 291(5507): 1260–1261.

Ross, P., Hall, L., Smirnov, I., and Haff, L. 1998. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nature Biotechnology* 16(13):1347–1351.

Rossmann, M. G., and Argos, P. 1976. Exploring structural homology of proteins. *Journal Molecular Biology* 105(1):75–95.

Roy, A., and Zhang, Y. 2012. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* 20(6):987–997.

Russell, R. B., and Barton, G. J. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Structure, Function, and Genetics* 14(2):309–323.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M., and Sali, A. 2004. A structural perspective on protein–protein interactions. *Current Opinion Structural Biology* 14(3):313–324.

Ryan, M., Diekhans, M., Lien, S., Liu, Y., and Karchin, R. 2009. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics* 25(11):1431–1432.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409(6822): 928–933.

Saibil, H. R. 2000. Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallographica Section D* 56(Pt 10):1215–1222.

Salem, S., Zaki, M. J., and Bystroff, C. 2010. FlexSnap: flexible non-sequential protein structure alignment. *Algorithms in Molecular Biology* 5(1):12.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* 32(Database issue):D449–D451.

Sambrook, J., and Russell, D. W. 2006. Identification of associated proteins by coimmunoprecipitation. *Cold Spring Harbor Protocols* 2006(1):pdb.prot3898.

Scaiewicz, A., and Levitt, M. 2015. The language of the protein universe. *Current Opinion in Genetics and Development* 35:50–56.

Schaefer, C., Meier, A., Rost, B., and Bromberg, Y. 2012. SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics* 28(4):601–602.

Schierz, A. C., Soldatova, L. N., and King, R. D. 2007. Overhauling the PDB. *Nature Biotechnology* 25(4):437–442.

Schmidt, B., Ho, L., and Hogg, P. J. 2006. Allosteric disulfide bonds. *Biochemistry* 45(24): 7429–7433.

Schofield, P. N., and Hancock, J. M. 2012. Integration of global resources for human genetic variation and disease. *Human Mutation* 33(5):813–816.

Schreyer, A., and Blundell, T. 2009. CREDO: a protein-ligand interaction database for drug discovery. *Chemical Biology and Drug Design* 73(2):157–167.

Schröder, G. F. 2015. Hybrid methods for macromolecular structure determination: experiment with expectations. *Current Opinion Structural Biology* 31:20–27.

Schuler, B., and Eaton, W. A. 2008. Protein folding studied by single-molecule FRET. *Current Opinion Structural Biology* 18(1):16–26.

Schuster-Böckler, B., and Bateman, A. 2008. Protein interactions in human genetic diseases. *Genome Biology* 9(1):R9.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. 2005. The FoldX web server: An online force field. *Nucleic Acids Research* 33(Web Server issue): W382–W388.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3):379–423.

Shatsky, M., Nussinov, R., and Wolfson, H. J. 2006. Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins* 62(1):209–217.

Shen, J., Deininger, P. L., and Zhao, H. 2006. Applications of computational algorithm tools to identify functional SNPs in cytokine genes. *Cytokine* 35(1-2):62–66.

Sherry, S. T., Ward, M., and Sirotkin, K. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research* 9(8): 677–679.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29(1):308–311.

Shi, D., Nannenga, B. L., Iadanza, M. G., and Gonen, T. 2013. Three-dimensional electron crystallography of protein microcrystals. *eLife* 2:e01345.

Shi, Y. 2014. A glimpse of structural biology through X-ray crystallography. *Cell* 159(5): 995–1014.

Shindyalov, I. N., and Bourne, P. E. 2001. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Research* 29(1):228–229.

Shoemaker, B. A., Zhang, D., Tyagi, M., Thangudu, R. R., Fong, J. H., Marchler-Bauer, A., Bryant, S. H., Madej, T., and Panchenko, A. R. 2012. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Research* 40(Database issue):D834–D840.

Shortle, D., Stites, W. E., and Meeker, A. K. 1990. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry* 29(35):8033–8041.

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., and Orengo, C. A. 2015. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research* 43(Database issue):D376–D381.

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R., Yeats, C., Thornton, J. M., and Orengo, C. A. 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Research* 41(Database issue): D490–D498.

Simonti, C. N., and Capra, J. A. 2015. The evolution of the human genome. *Current Opinion in Genetics and Development* 35:9–15.

Soon, W. W., Hariharan, M., and Snyder, M. P. 2013. High-throughput sequencing for biology and medicine. *Molecular Systems Biology* 9(1):640–640.

Sowdhamini, R., Burke, D. F., Huang, J. F., Mizuguchi, K., Nagarajaram, H. A., Srinivasan, N., Steward, R. E., and Blundell, T. L. 1998. CAMPASS: a database of structurally aligned protein superfamilies. *Structure* 6(9):1087–1094.

Sprinzak, E., Sattath, S., and Margalit, H. 2003. How reliable are experimental protein-protein interaction data? *Journal Molecular Biology* 327(5):919–923.

Sreerama, N., Venyaminov, S., and Woody, R. W. 1999. Estimation of the number of α-helical and β-strand segments in proteins using circular dichroism spectroscopy. *Protein Science*.

Stankiewicz, P., and Lupski, J. R. 2010. Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine* 61(1):437–455.

Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., and Alexov, E. 2013. Molecular mechanisms of disease-causing missense mutations. *Journal Molecular Biology* 425(21): 3919–3936.

Stein, A. 2004. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Research* 33(Database issue):D413–D417.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., et al. 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122(6):957–968.

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., and Cooper, D. N. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics* 133(1):1–9.

Steward, R. E., MacArthur, M. W., Laskowski, R. A., and Thornton, J. M. 2003. Molecular basis of inherited diseases: a structural perspective. *Trends in Genetics* 19(9):505–513.

Stitziel, N. O., Tseng, Y. Y., Pervouchine, D., Goddeau, D., Kasif, S., and Liang, J. 2003. Structural location of disease-associated single-nucleotide polymorphisms. *Journal Molecular Biology* 327(5):1021–1030.

Stone, E. A., and Sidow, A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research* 15(7):978–986.

Strickler, S. S., Gribenko, A. V., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V., and Makhatadze, G. I. 2006. Protein Stability and Surface Electrostatics: A Charged Relationship. *Biochemistry* 45(9):2761–2766.

Sunyaev, S., Lathe, W., and Bork, P. 2001. Integration of genome data and protein structures: prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms. *Current Opinion Structural Biology* 11(1):125–130.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156(6):1324–1335.

Syvänen, A. C. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* 2(12):930–942.

Tang, H., Wyckoff, G. J., Lu, J., and Wu, C.-I. 2004. A universal evolutionary index for amino acid changes. *Molecular Biology and Evolution* 21(8):1548–1556.

Taylor, W. R. 1986. The classification of amino acid conservation. *Journal of Theoretical Biology* 119(2):205–218.

Taylor, W. R. 2007. Evolutionary transitions in protein fold space. *Current Opinion Structural Biology* 17(3):354–361.

Teng, S., Madej, T., Panchenko, A., and Alexov, E. 2009. Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. *Biophysics Journal* 96(6):2178–2188.

Teng, S., Srivastava, A. K., Schwartz, C. E., Alexov, E., and Wang, L. 2010. Structural assessment of the effects of amino acid substitutions on protein stability and protein protein interaction. *International Journal of Computational Biology and Drug Design* 3(4):334–349.

Terp, B. N., Cooper, D. N., Christensen, I. T., Jørgensen, F. S., Bross, P., Gregersen, N., and Krawczak, M. 2002. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. *Human Mutation* 20(2):98–109.

Teyra, J., Samsonov, S. A., Schreiber, S., and Pisabarro, M. T. 2011. SCOWLP update: 3D classification of protein-protein, -peptide, -saccharide and -nucleic acid interactions, and structure-based binding inferences across folds. *BMC Bioinformatics* 12(1):398.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861.

The UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research* 42(Database issue):D191–D198.

The UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Research* 43(Database issue):D204–D212.

Thompson, J. D., Plewniak, F., and Poch, O. 1999. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* 15(1):87–88.

Thornton, J. M., Orengo, C. A., Todd, A. E., and Pearl, F. M. 1999. Protein folds, functions and evolution. *Journal Molecular Biology* 293(2):333–342.

Thusberg, J., Olatubosun, A., and Vihinen, M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation* 32(4):358–368.

Tinti, M., Dissanayake, K., Synowsky, S., Albergante, L., and MacKintosh, C. 2014. Identification of 2R-ohnologue gene families displaying the same mutation-load skew in multiple cancers. *Open Biology* 4(5):140029–140029.

Todd, A. E., Orengo, C. A., and Thornton, J. M. 1999. Evolution of protein function, from a structural perspective. *Current Opinion in Chemical Biology* 3(5):548–556.

Tsai, J., Taylor, R., Chothia, C., and Gerstein, M. 1999. The packing density in proteins: standard radii and volumes. *Journal Molecular Biology* 290(1):253–266.

Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. 2011. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature Protocols* 6(9):1341–1354.

Valdar, W. S., and Thornton, J. M. 2001. Conservation helps to identify biologically relevant crystal contacts. *Journal Molecular Biology* 313(2):399–416.

Valdar, W. S. J. 2002. Scoring residue conservation. *Proteins: Structure, Function, and Genetics* 48(2):227–241.

Vázquez, M., Valencia, A., and Pons, T. 2015. Structure-PPi: A module for the annotation of cancer-related single-nucleotide variants at protein-protein interfaces. *Bioinformatics* 31(14):2397–2399.

Velankar, S., Best, C., Beuth, B., Boutselakis, C. H., Cobley, N., Sousa Da Silva, A. W., Dimitropoulos, D., Golovin, A., Hirshberg, M., John, M., Krissinel, E. B., Newman, R., Oldfield, T., Pajon, A., Penkett, C. J., et al. 2010. PDBe: Protein Data Bank in Europe. *Nucleic Acids Research* 38(Database issue):D308–D317.

Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research* 41(Database issue):483–489.

Velankar, S., van Ginkel, G., Alhroub, Y., Battle, G. M., Berrisford, J. M., Conroy, M. J., Dana, J. M., Gore, S. P., Gutmanas, A., Haslam, P., Hendrickx, P. M. S., Lagerstedt, I., Mir, S., Fernandez Montecelo, M. A., Mukhopadhyay, A., et al. 2016. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Research* 44(Database issue):D385–D395.

Veretnik, S., Bourne, P. E., Alexandrov, N. N., and Shindyalov, I. N. 2004. Toward consistent assignment of structural domains in proteins. *Journal Molecular Biology* 339(3):647–678.

Visel, A., Zhu, Y., May, D., Afzal, V., Gong, E., Attanasio, C., Blow, M. J., Cohen, J. C., Rubin, E. M., and Pennacchio, L. A. 2010. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* 464(7287):409–412.

Vitkup, D., Sander, C., and Church, G. M. 2003. The amino-acid mutational spectrum of human genetic disease. *Genome Biology* 4(11):R72.

Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., et al. 2013. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research* 41(Database issue):D1063–D1069.

Voet, D., and Voet, J. G. 2010. *Biochemistry, 4th Edition.* John Wiley & Sons.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. 2004. Structure, function and evolution of multidomain proteins. *Current Opinion Structural Biology* 14(2):208–216.

Wang, Z., and Moult, J. 2001. SNPs, protein structure, and disease. *Human Mutation* 17(4):263–270.

Wang, Z., and Moult, J. 2003. Three-Dimensional Structural Location and Molecular Functional Effects of Missense SNPs in the T Cell Receptor Vβ Domain. *Proteins: Structure, Function, and Genetics* 53(3):748–757.

Ward, L. D., and Kellis, M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology* 30(11):1095–1106.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. 2009. Jalview Version 2: a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145):661–678.

Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W., Weissig, H., Greer, D. S., Bourne, P. E., and Berman, H. M. 2002. The Protein Data Bank: unifying the archive. *Nucleic Acids Research* 30(1):245–248.

Westbrook, J. 2012. PDBx Python Module. `http://mmcif.wwpdb.org/docs/sw-examples/python/html/index.html`. Accessed: 2014-01-12.

Winter, C. 2006. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Research* 34(90001):D310–D314.

Wlodawer, A., Minor, W., Dauter, Z., and Jaskolski, M. 2007. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS Journal* 275(1):1–21.

Worth, C. L., Gong, S., and Blundell, T. L. 2009. Structural and functional constraints in the evolution of protein families. *Nature Reviews Molecular Cell Biology* 10(10): 709–720.

Worth, C. L., Preissner, R., and Blundell, T. L. 2011. SDM–a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Research* 39(Web Server issue):W215–W222.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N., et al. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* 34(Database issue): D187–D191.

Xia, Y., and Levitt, M. 2004. Simulating protein evolution in sequence and structure space. *Current Opinion Structural Biology* 14(2):202–207.

Xie, L., and Bourns, P. E. 2005. Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Computational Biology* 1(3):222–229.

Yang, J., Roy, A., and Zhang, Y. 2012. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research* 41(Database issue):D1096–D1103.

Yang, J. O., Oh, S., Ko, G., Park, S. J., Kim, W. Y., Lee, B., and Lee, S. 2011. VnD: A structure-centric database of disease-related SNPs and drugs. *Nucleic Acids Research* 39(Database issue):D939–D944.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. S., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P. 2014. The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics* 31(1):143–145.

Yates, C. M., and Sternberg, M. J. E. 2013a. Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms (nsSNPs). *Journal Molecular Biology* 425(8):1274–1286.

Yates, C. M., and Sternberg, M. J. E. 2013b. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *Journal Molecular Biology* 425(21):3949–3963.

Ye, Y., and Godzik, A. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. In *Bioinformatics*, ii246–255. Sanford Burnham Prebys Medical Discovery Institute, San Diego CA, United States.

Yu, B., Sawyer, N. A., Chiu, C., Oefner, P. J., and Underhill, P. A. 2001. *Dna mutation detection using denaturing high-performance liquid chromatography (dhplc)*. John Wiley & Sons.

Yue, P., Melamud, E., and Moult, J. 2006. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 7(1):166.

Yue, P., and Moult, J. 2006. Identification and analysis of deleterious human SNPs. *Journal Molecular Biology* 356(5):1263–1274.

Zajonc, D. M., Elsliger, M. A., Teyton, L., and Wilson, I. A. 2003. Crystal structure of CD1a in complex with a sulfatide self antigen at a resolution of 2.15 A. *Nature Immunology* 4(8):808–815.

Zamyatnin, A. A. 1972. Protein volume in solution. *Progress in Biophysics and Molecular Biology* 24(C):107–123.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., and Honig, B. 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 490(7421):556–560.

Zhang, Y., and Skolnick, J. 2005. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* 33(7):2302–2309.

Zhang, Z., Wang, Y., Wang, L., and Gao, P. 2010. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS ONE* 5(12):e14316.

Zhanhua, C., Gan, J. G.-K., Lei, L., Sakharkar, M. K., and Kangueane, P. 2005. Protein subunit interfaces: heterodimers versus homodimers. *Bioinformation* 1(2):28–39.

Zhao, C., and Sacan, A. 2015. UniAlign: protein structure alignment meets evolution. *Bioinformatics* 31(19):3139–3146.

Zhao, N., Han, J. G., Shyu, C.-R., and Korkin, D. 2014. Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. *PLoS Computational Biology* 10(5):e1003592.